IDENTIFICAÇÃO DE PADRÕES EM INFRAESTRUTURA FERROVIÁRIA – VIA

UMA ABORDAGEM *DATA MINING*

por

Nuno Pinto Barriga de Carvalho Tavares

Tese de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientada por:
Prof. Dr. João Gama
Prof. Dr. Alípio Jorge

Faculdade de Economia
Universidade do Porto
2014

Nota Biográfica

Nuno Pinto Barriga de Carvalho Tavares licenciou-se em Economia pela Universidade Lusíada de Lisboa em 1998, sendo pós-graduado em Economia de Gestão em 2000 pela mesma universidade, especializado em Controlo de Gestão em 2005 pelo INDEG – ISCTE.

Profissionalmente ocupou vários cargos no domínio dos sistemas de apoio à decisão e controlo de gestão, tais como, Diretor do Departamento de Controlo de Gestão da Delegação Norte da REFER, responsável pelo Departamento de Controlo de Gestão e Planeamento da Direção de Investimentos da REFER.

Paralelamente frequentou vários cursos no âmbito profissional ligados às áreas de apoio à decisão, nos domínios do planeamento e controlo de gestão. Foi ainda orador na IV Conferência do *Project Management Institute* (PMI), sendo um entusiasta dos sistemas de apoio à decisão, controlo de gestão e planeamento.

Agradecimentos

À Susete sem a qual nada teria sido possível, aos meus filhos, Pedro e Maria, pela generosidade e compreensão face ao tempo que não partilhámos.

À minha família, à minha mãe e ao meu irmão, e a todos os meus amigos que nos últimos 2 anos se viram privados da minha companhia.

Aos meus orientadores Prof. João Gama e Prof. Alípio Jorge, pela sábia orientação e disponibilidade permanente.

Este trabalho foi suportado pelo projeto de investigação Sibila (NORTE-07-0124-FEDER-000059), financiado pelo Programa Operacional da Região Norte Portugal (ON.2 O Novo Norte), sob o quadro estratégico de referência (NSRF), pelo Fundo de Desenvolvimento (ERDF), e por fundos nacionais, através da agência Portuguesa, Fundação para a Ciência e a Tecnologia (FCT), e pela Comissão Europeia através do projeto MAESTRA (Grant number ICT-2013-612944).

Resumo

Os gestores de infraestruturas dispõem de um número crescente de informação em bases de dados resultantes dos processos de inspeção automatizados. Estas constituem uma fonte de conhecimento que aguarda ser explorada, pois a abordagem tradicional à aquisição de conhecimento nestas áreas tem versado sobre a modelação estatística associada aos fenómenos de degradação, que explora a base de dados com uma orientação prévia guiada pelos paradigmas de engenharia instituídos.

O presente trabalho efetua, sob a forma de um estudo de caso, uma proposta de abordagem, num contexto de aprendizagem não supervisionada, guiada pelos dados, sem prévios constrangimentos. Este culmina na aplicação de um algoritmo, denominadas regras de distribuição, com um elevado grau de completude, quando por exemplo comparado com as técnicas de regressão tradicionalmente utilizadas nesta área. A interpretabilidade é também uma das suas caraterísticas, diferenciando-o de outras metodologias como por exemplo redes neuronais.

Os principais resultados prendem-se com a capacidade de ter gerado mais de 600 padrões considerados estatisticamente diferentes das distribuições à priori, efetuado uma proposta de apresentação dos mesmos ao utilizador num único suporte facilmente interpretável, e ter proposto uma abordagem à classificação do interesse desses padrões objetiva e subjetivamente.

Os impactos deste trabalho serão o de contribuir para acelerar a introdução das técnicas *de Data Mining* na gestão de infraestruturas, nomeadamente ferroviárias, como meio de extração de conhecimento e suporte à decisão, bem como a introdução da presente abordagem e respetivo algoritmo na empresa subjacente ao estudo de caso.

Palavras-chave: Regras de Distribuição, Gestão de Infraestruturas - *Data Mining*.

Índice

Índice de Figuras

Índice de Tabelas

Índice de Expressões

1. Introdução

A disponibilidade de milhões de registos sobre o estado das infraestruturas provenientes das inspeções efetuadas é uma oportunidade para a extração de conhecimento que permitirá um melhor conhecimento e gestão das mesmas. Esta necessidade é sublinhada pelo atual quadro de eficiência que é hoje exigido aos gestores de infraestruturas, conforme Plano Estratégico de Transportes estabelecido pelo Ministério da Economia e do Emprego [1].

É pois objetivo do presente trabalho a aplicação de um algoritmo que permita a extração de conhecimento dos registos existentes, sendo eficiente computacionalmente, interpretável e completo, por oposição às técnicas de regressão habitualmente utilizadas nesta indústria, que apenas determinam uma relação do fenómeno em estudo. O algoritmo deve ser complementado com um interface que possibilite ao utilizador/especialista de domínio encontrar padrões interessantes com o mínimo dispêndio de tempo.

O que se procura é que o especialista de domínio tenha acesso às causas de degradação da infraestrutura, presentes nas bases de dados sobre a forma de caraterísticas/componentes, com o intuito de, antevendo o seu comportamento elaborar planos eficientes de manutenção e renovação.

De forma a corresponder aos requisitos enunciados é implementado um algoritmo baseado num modelo apresentado por Jorge, Azevedo and Pereira [2], denominado "Distribution rules with numeric attributes of interest". Este modelo visa identificar padrões frequentes de associação, subgrupos, e efetuar previsões, de uma forma similar ao modelo de regras de associação Agrawal R [3], com a diferença que associa e identifica uma distribuição no consequente.

Um segundo objetivo é identificar e propor à organização um método de *Data Mining*, seguindo a metodologia CRISP-DM Chapman [4], que permita replicar o processo. O tratamento periódico da base de dados utilizada potenciará a extração de conhecimento.

Este irá ser mais facilmente usado no plano operacional da organização para que esta possa mais eficiente e eficazmente atingir os seus objetivos.

Desta forma a estrutura do documento irá ter como referência as fases propostas na metodologia CRISP-DM Chapman [4], que se auto define como uma metodologia *Step By Step.* A abordagem efetuada pela metodologia imprime uma forte relação entre os objetivos da empresa e a ação do *Data Mining.* O trabalho em apreço tentou adequar o corpo normal de uma tese a estas duas características, efetuando um enquadramento mais extenso sobre o negócio, adotando uma perspetiva funcional dos conceitos subjacentes ao corpo tradicional de uma tese. Como seja integrar o ponto revisão da literatura no ponto 2. "Avaliação da Situação", no ponto 7 "Modelação do Negócio" e documentar de forma detalhada os passos efetuados.

Uma terceira caraterística da metodologia é não ter uma estrutura sequencial, os pontos definidos podem ser recursivos, o que se passou também nesta tese pois os dados obrigaram à repetição da abordagem para os pontos análise, qualidade e preparação dos dados.
Desta forma esta tese encontra-se estruturada em 8 capítulos mais a presente introdução da seguinte forma:

1. Introdução.

2. Compreender o Negócio.
   Neste ponto é objetivo dar a conhecer o contexto de atuação do trabalho no âmbito do negócio, realçando os seus objetivos, efetuando a avaliação dos recursos existentes, riscos, constrangimentos, entre outros pontos, numa abordagem de gestão de um projeto.

3. Conhecimento dos Dados.
   Efetua-se a análise aos dados e define-se a estratégia do seu tratamento.

4. Preparação dos Dados.

5. Integração dos Dados.

6. Análise Qualidade e Preparação dos Dados.
   A recursividade referida manifestou-se neste ponto, que é a repetição dos subpontos 3.3., 3.4. e 4., em função de se ter tratado os dados em dois momentos por impossibilidade da integração inicial dos ficheiros.

7. Modelação.
   Escolha, modelação e avaliação do algoritmo.

8. Avaliação.
   Avaliação segundo a perspetiva do negócio. Neste ponto recorreu-se aos conceitos de interesse da regra para tentar mensurar o sucesso do algoritmo à luz do conceito de negócio.

9. Conclusões.
   Ponto não pertencente à metodologia que visa retirar as conclusões do trabalho subjacente a esta tese.

Cumpre referir, para os leitores mais próximos da área, ligados à especialidade de Via de infraestruturas ferroviárias, que este documento é tão só uma proposta de metodologia para a produção de conhecimento no âmbito do *framework* do *Data Mining*, acompanhando a mudança de paradigma associada à análise da temática em questão provocada pelo incremento das bases de dados e processos de inspeção/leitura automatizados.

2. Compreender o Negócio

2.1. Objetivos do Negócio

Uma empresa de gestão de infraestruturas ferroviárias tem por objetivo a disponibilização da infraestrutura[1] com um nível de serviço pré definido ao menor custo possível.

Conforme disposto por Ministério da Economia e do Emprego [1] no Plano Estratégico de Transportes ponto 4.5.2.1. Racionalização de Custos, as empresas do Setor Empresarial do Estado irão tomar as medidas necessárias à redução em 15%, face a 2009, dos custos com fornecimentos e serviços externos e outros custos operacionais.

Tendo em conta que a maior parcela de origem de custos deste tipo de empresas tem origem na razão de ser do seu negócio, a manutenção da infraestrutura, a compreensão do processo a ela associado, relativo à degradação da infraestrutura, é central a qualquer tentativa de redução de custos.

Deste modo o desconhecimento relativo sobre o processo de degradação da infraestrutura, é um dos fatores principais de origem de custos neste tipo de empresas, sendo que as organizações pagam prémios elevados, quer internamente, na medida em que não permite um alinhamento completo dos seus recursos internos, quer em termos externos, em sede de contratação, em que paga um preço suplementar para suportar a imprevisibilidade, que é perfeitamente documentado e objeto de articulado contratual próprio na atividade de manutenção[2].

---

[1] Por infraestrutura entende-se toda a estrutura que se encontra diretamente ligada ao objeto do negócio, como linhas (e todas as suas componentes sinalização, catenária, via), obras de arte (pontes e túneis), estações e outras estruturas.

[2] Por regra os contratos de manutenção distinguem dois tipos de preço para o mesmo âmbito de trabalho com base na maior ou menor rapidez de intervenção.

Este desconhecimento relativo assenta na complexidade que apresenta o fenómeno de degradação da infraestrutura e subsequente previsão de intervenção, resultante de um conjunto alargado de variáveis que, potencialmente, intervêm no estado da infraestrutura.

A infraestrutura ferroviária subdivide-se nas seguintes especialidades:

- Via (especialidade que será o nosso objeto de estudo);
- Catenária;
- Sinalização.

Destas, só a especialidade de via será objeto de estudo, decompondo-se da seguinte forma:

- Infraestrutura de Via (camadas localizadas sob o balastro):
  - Aterros;
  - Taludes de Escavação;
  - Sistemas de Drenagem;
  - Obras de Arte (destinadas a suportar a via).

- Superestrutura de Via:
  - Travessas (Betão (várias modalidades), Madeira (várias modalidades));
  - Carril (várias modalidades e dois tipos de ligação entre carris);
  - Fixações (diversos tipos);
  - Ligações;
  - Balastro (camada de inertes com propriedades que absorvem os impactos resultantes das passagens das composições (comboios)).

A acrescer à diversidade acima identificada, existe um conjunto de variáveis que contextualizam a infraestrutura tais como:

- Tráfego - N.º de comboios, comprimento médio, peso total (MGT[3]), velocidade máxima;
- Condições Meteorológicas – Temperatura média, amplitude térmica, humidade, pluviosidade;
- A altimetria da via;
- A planimetria da via.

São todas estas variáveis em interação, com a multiplicidade de valores que cada uma apresenta, que qualificam a complexidade de previsão da degradação e subsequente plano de manutenção.

Não sendo do conhecimento do autor uma atividade corrente na área de *Data Mining* na empresa em apreço, sendo necessário a diminuição dos custos operacionais, existindo milhões de registos sobre o comportamento da infraestrutura, é um objetivo natural implementar um processo que periodicamente retire conhecimento das bases de dados existentes no sentido de compreender a relação das variáveis mencionadas e a degradação da infraestrutura, por forma a obter:

- Uma atuação mais preventiva na infraestrutura;
- Um maior cumprimento dos planos de manutenção;
- Melhor disciplina na afetação recursos externos e internos;
- Intervenções de renovação e criação de nova infraestrutura com componentes com melhor eficiência na relação custo/desempenho, medido pela degradação de valores associada ao seu ciclo de vida (exº. fadiga do material por nº de toneladas suportada).

Cumpre referir, conforme será melhor explicitado no capítulo 8, que no contexto do algoritmo proposto, regras de distribuição, o conhecimento é materializado como um ou

[3] Acrónimo da expressão inglesa Million Gross Ton.

vários padrões interessantes pelo que a descoberta dos mesmos diminui a imprevisibilidade do negócio, promovendo os objetivos supra.

Conforme metodologia CRISP-DM Chapman [4] a avaliação do sucesso deste trabalho, do ponto de vista do "negócio", deveria estabelecer para cada um dos objetivos acima definidos um critério de avaliação, contudo face à ausência de uma métrica que permitisse inequivocamente definir uma relação entre a descoberta de um padrão interessante e cada um dos objetivos supra define-se que:

- A obtenção de um padrão considerado interessante por um técnico da área/especialista de domínio é critério de sucesso.

2.2. Avaliação da Situação

Segundo a metodologia CRISP-DM Chapman [4] este ponto efetua uma avaliação da situação com base numa inventariação de recursos disponíveis para o projeto como sejam: recursos de hardware, fontes de informação e conhecimento, recursos humanos. Identifica também os requisitos do projeto, pressupostos e constrangimentos, os riscos e contingências, terminologia e custos e benefícios.

A adaptação da metodologia ao trabalho em curso levou à alteração de alguns pontos, como "fontes de informação e conhecimento" decompõe-se no presente trabalho em "2.2.1. Informação Disponível na empresa Bases de Dados e Fontes de informação" e "2.2.2. Revisão da Literatura " e a supressão de outros pontos para os quais não existiriam meios para o cálculo dos mesmos como seja "Custos e Benefícios".

A infraestrutura tecnológica de suporte ao apoio à decisão dos gestores de infraestruturas ferroviários, no contexto da infraestrutura de via, encontra-se mobilizada para responder à garantia da qualidade da infraestrutura, através da classificação da informação decorrente das inspeções efetuadas e monitorização dos seus valores no quadro de referências definidos pelo normativo IT.VIA.018 REFER [5] que é extensão da norma Europeia para o sector EN 13848-5 (CEN) [6].

Conforme Andrade [7], de acordo com a tabela 1, os defeitos de via são habitualmente monitorizados através de sete parâmetros:

Defeitos de Via	FE*	FD*
Nivelamento Longitudinal	x	x
Alinhamento	x	x
Nivelamento Transversal	-	-
Bitola	-	-
Empeno	-	-

* Fila de carril esquerda/direita

Tabela 1: Parâmetros de defeitos de via

Estes são posteriormente "transformados em indicadores" em que se avalia:

- O seu valor pontual;
- O desvio padrão;
- A sua média.

Esta avaliação é efetuada através da classificação dos valores destes indicadores numa grelha constante da norma EN 13848-5 (CEN) [6], que na sua extensão portuguesa IT.VIA .018 REFER [5] define, face aos valores em presença, as ações a implementar como:

- Ação Imediata- São valores definidos como imperativos em sede de EN 13848-5 (CEN) [6], não sujeitos a interpretação pelo gestores das infraestruturas, pois

traduzem uma relação direta entre a roda (da composição ferroviária) e o carril, proveniente da experiência e teoria produzida sobre a matéria. Quando excedidos obrigam à redução da velocidade máxima da circulação dos comboios ou em casos excecionais à suspensão do troço/linha em que se registaram os valores.

- Intervenção – Valores que, se excedidos, requerem uma intervenção corretiva para que não alcancem o limite de segurança (definido como intervenção imediata) até à próxima intervenção.

- Alerta - Valores que, se excedidos, requerem análise às condições de geometria da via sendo as ações tendentes à correção inseridas em sede de planeamento das ações de manutenção a efetuar.

Conforme referido os limites de ação imediata são definidos em sede de EN 13848-5 (CEN) [6] sendo considerados imperativos. Os limites de intervenção e de alerta são definidos por cada um dos gestores de infraestrutura atendendo à sua política de manutenção.

Segundo Andrade [7] esta política de manutenção assenta em quatro dimensões calibradas pela maior ou menor exigência com o nível de serviço definido pelos gestores de infraestruturas, através da definição dos valores limite para parâmetros geométricos de via, segurança, conforto, custo do ciclo de vida do ativo (troço, aparelho de mudança de via, outras componentes da infraestrutura de via) e disponibilidade do canal.

A velocidade é um fator que se correlaciona positivamente com a qualidade de via, sendo que, conforme tabela 2 quanto maior a velocidade mais rigorosos são os valores de referência para a qualidade dos parâmetros de via.

A tabela 2, 3 e 4 (infra) retiradas da EN 13848-5 (CEN) [6] são exemplo das diversas formas de aferição dos parâmetros. No que respeita à tabela 2 encontramos a aferição

para o parâmetro geométrico de via bitola, atendendo ao defeito isolado ou valor pontual.

Speed (in km/h)	Nominal track gauge to peak value (in mm) *IAL*	
	Minimum	Maximum
$V \leq 80$	-11	+35
$80 < V \leq 120$	-11	+35
$120 < V \leq 160$	-10	+35
$160 < V \leq 230$	-7	+28
$230 < V \leq 300$	-5	+28

Tabela 2: Velocidade e os parâmetros geométricos de via

A tabela 3 apresenta os valores em função do desvio do valor pontual para a média dos valores da bitola em 100 metros.

Table 3 — Track gauge – *IAL* – Nominal track gauge to mean track gauge over 100 m

Speed (in km/h)	Nominal track gauge to mean track gauge over 100 m (in mm)	
	Minimum	Maximum
$V \leq 40$	*N/A*	+32
$40 < V \leq 80$	-8	+32
$80 < V \leq 120$	-7	+27
$120 < V \leq 160$	-5	+20
$160 < V \leq 230$	-5	+20
$230 < V \leq 300$	-3	+20

Tabela 3: Parâmetros geométricos de via média

A tabela 4 apresenta os valores do nivelamento longitudinal calculados para o desvio padrão em comprimentos de onda [4] de 3 a 25 metros (D1) e 27 a 70 metros (D2)

[4] Comprimento de onda é a distancia entre valores repetidos num padrão de onda, e é utilizado para aferir padrões que possam provocar forças "anómalas" na interação roda carril . A prática corrente adotou o comprimento de onda 3-

repartidos pelos três níveis de ação de crescente necessidade de intervenção, Alerta (AL), Intervenção (IL) e Ação Imediata (IAL).

B.2.2 Longitudinal level

Table B.3 — Longitudinal level – *AL* & *IL* – Isolated defects – Mean to peak value

Speed (V in km/h)	Mean to peak value (a_0 in mm) *AL*		Mean to peak value (a_0 in mm) *IL*		Mean to peak value (a_0 in mm) *IAL (residual)*	
	Wavelength range		Wavelength range		Wavelength range	
	D1	*D2*	*D1*	*D2*	*D1*	*D2*
$V \leq 80$	12 to 18	*N/A*	17 to 21	*N/A*	28	*N/A*
$80 < V \leq 120$	10 to 16	*N/A*	13 to 19	*N/A*	28	*N/A*
$120 < V \leq 160$	8 to 15	*N/A*	10 to 17	*N/A*	23	*N/A*
$160 < V \leq 230$	7 to 12	14 to 20	9 to 14	18 to 23	20	33
$230 < V \leq 300$	6 to 10	12 to 18	8 to 12	16 to 20	16	28

Tabela 4: Parâmetros geométricos de via em comprimento de onda D1 e D2

Nesta "Avaliação de Situação" expôs-se como atualmente os gestores de infraestruturas utilizam os dados, sendo sobretudo para correção e identificação de falhas e alertas.

O presente trabalho visa atender aos dados para perceber o que originam essas falhas, por forma a se prever no futuro o comportamento da infraestrutura.

2.2.1. Informação Disponível na Empresa Bases de Dados e Fontes de informação.

Bases de Dados

Conforme melhor identificado no Apêndice I as bases de dados existentes e relevantes para o presente trabalho são o ficheiro G_Medições contendo as medições dos parâmetros geométricos de via resultantes das inspeções e a base de dados T_Tables

---

25 metros como padrão, nomeadamente para o nivelamento longitudinal e alinhamento que são os que provocam as forças verticais e horizontais associadas à segurança das composições e conforto dos passageiros, para a aferição da qualidade dos parâmetros medidos pelo desvio padrão. Conforme guia das melhores práticas em otimização da durabilidade de geometria de via [8] UIC *Best practice guide for optimum track geometry durability.* City, 2008. o desvio padrão para a onda de cumprimento (3-25m) dos defeitos longitudinais é o indicador crucial para as decisões de manutenção.

relativa às caraterísticas da circulação (veículos ferroviários) que utilizaram a infraestrutura.

Outras Fontes de Dados

Foram também identificadas várias fontes de dados em suporte físico e suporte informático, mas não no formato de base de dados[5], com informação relativamente às componentes da infraestrutura.

Outras fontes de informação, já mencionadas, prendem-se com os normativos internos nomeadamente IT.VIA.018 [10] que classificam o valor dos parâmetros em termos da sua aceitabilidade e "urgência" de intervenção retificativa.

2.2.2. Revisão da Literatura

Abordagem "Corrente"

Conforme Andrade [7] a abordagem mais comum na literatura à questão da determinação da origem das falhas dos componentes de infraestrutura (carril, travessas, fixações) tem sido:

- A modelação estatística centrada na identificação das distribuições associadas aos fenómenos de degradação e respetivos elementos caraterizadores (quer de centralidade quer de dispersão), nomeadamente por comparação dos desvios padrões dos parâmetros identificados na tabela 1;
- A elaboração de modelos de degradação através de técnicas de regressão.

[5] Conforme [9] João Gama, A. C. L., Katti Faceli, Andre Ponce de Leon Carvalho,Márcia Oliveira *Extração de Conhecimento de Dados*. Edições Sílabo, lda, 2012. "Formalmente um conjunto de dados pode ser representado por uma matriz de objetos X_nXd em que n é o número de objetos e d o número de variáveis associadas a cada objeto.

Esta abordagem coloca um especial enfoque no nivelamento longitudinal e alinhamento, agrupados por segmentos de via e classificados por diferentes tipologias que de acordo com Andrade [7] assumem-se como atributos ”explicadores” da degradação da via e parâmetros correspondentes.

Abordagem *Data Mining*

A abordagem proposta neste trabalho baseia-se no tratamento dos dados em bruto, através do framework fornecido pelo *Data Mining*, sem quaisquer pressupostos e ou condicionantes para entender quais os padrões existentes que os dados revelam.

Esta abordagem envolve algum risco pois não aproveita à partida um conhecimento já existente. Contudo também se suporta na visão de que a complexidade de uma infraestrutura de via vai para além dos atributos habitualmente identificados como explicativos do processo de degradação a saber, conforme López Pita [11] seções de 200 metros que contém ou se situam em plena via, estações, pontes ou e AMV´S.

Assim, conforme Esveld [12] os componentes de via têm supostamente propriedades mecânicas especificas que permitem à infraestrutura suportar e guiar as composições ferroviárias, definindo as suas propriedades conforme figura infra:

	Elasticity	Strength	Stability	Durability
Rail profile		X		X
Fastening system	X		X	X
Sleepers		X	X	X
Ballast	X		X	X
Slabs		X		X
Track support systems	X		X	X

Tabela 5: Propriedades da via mais relevantes por componente

Desta forma pretende-se explorar um método que possibilite sem restrições relacionar o maior número de fatores relacionados com a via e a sua performance. Esta visão face às bases de dados existentes com problemas de fidedignidade e até ausência de informação

pode ficar à partida bastante comprometida, contudo se atendermos à evolução da informação registada em bases de dados revela-se bastante promissora. Conforme João Gama [9] estima-se que a cada 20 meses a quantidade de dados armazenada em todas as bases de dados do mundo duplica.

A revisão da literatura atende a esta visão de "*Data Mining*" , tentando identificar o que no âmbito do *framework* deste corpo de saber foi efetuado à data para infraestruturas. Para uma leitura relativa à bibliografia assente na modelação estatística, para identificação dos padrões associados à degradação da estrutura de via recomenda-se Andrade [7] capítulo 2.3. .

O problema de diagnóstico e previsão de degradação de infraestruturas, tem sido abordado na área do *Machine Learning* (modelos supervisionados) e ou *Data Mining* (modelos não supervisionados) de forma diversa em várias publicações:

Na área do fornecimento elétrico Gross, Boulanger, Arias, Waltz, Long, Lawson, Anderson, Koenig, Mastrocinque and Fairechio [13], em que utilizam um modelo de Boosting, Martingale Boosting Long and Servedio [14] que, combinando vários classificadores "fracos", funções definidas através de rankings produzidos pelos atributos, obtém um ranking de estruturas a intervencionar, interpretável para os utilizadores.

Este modelo possibilita, também, o diagnóstico, ao contrário de outros modelos de ranking referidos nesta publicação, como sejam o RankBoost Freund, Iyer, Schapire and Singer [10] e o Rank Net Burges, Shaked, Renshaw, Lazier, Deeds, Hamilton and Hullender [15], denominados pelos autores de caixas pretas. O primeiro utiliza um modelo de *boosting* e o segundo uma rede neuronal. Este modelo encontra-se aplicado por uma empresa de distribuição elétrica de Nova York, possibilitando direcionar a manutenção da rede de distribuição elétrica para onde irá ocorrer a falha.

Na área da infraestrutura ferroviária existem diversas publicações que, com diferentes abordagens, interpretam a predição e diagnóstico de falha. Da pesquisa efetuada são identificados, normalmente, dois subtemas associadas ao tema desta tese:

- O subtema da recolha dos dados – que, face à extensão e periodicidade de leituras associadas à infraestrutura ferroviária, tem motivado diversas publicações sobre o tema automação da inspeção.
- O subtema do tratamento dos dados – em que utilizam técnicas de *Machine Learning*, para tratarem problemas relacionados com a gestão do conhecimento/apoio à decisão na infraestrutura ferroviária - Via.

Subtema Recolha de Dados

Como exemplo, temos a publicação efetuada por Molina, Resendiz, Edwards, Hart, Barkan and Ahuja [16] que efetua um resumo dos sistemas e métodos associados ao que denomina de *Machine Vision*.

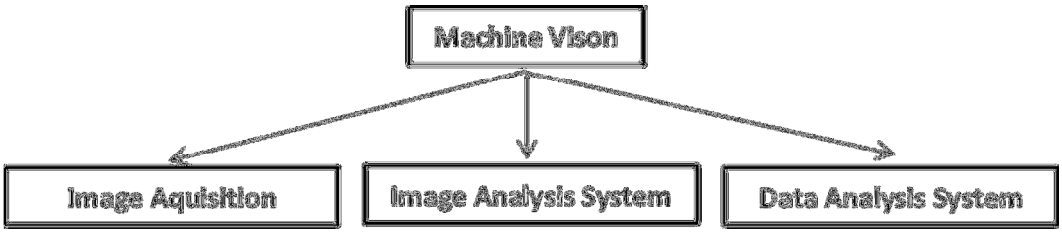


Figura 1: *Machine Vison* segundo Molina *et al.* (2010)

Conforme Figura 1, Molina, Resendiz, Edwards, Hart, Barkan and Ahuja [16] decompõem em três grandes elementos a *Machine Vision,* sendo que todos eles concorrem para o objetivo de identificar, de forma automática, falhas na infraestrutura ferroviária de via (Via), socorrendo-se de diversos sistemas de captação de imagem e de algoritmos de interpretação e classificação de imagens. A vantagem destes sistemas, do ponto de vista do *Machine Learning* e ou *Data Mining*, é tornar objetiva e sistematizada o processo de recolha de dados associado aos parâmetros de Via.

Resendiz, Hart and Ahuja [17] abordam este tema, apresentando três algoritmos, um para deteção de componentes de via (carril, travessa e balastro), em que recorre a máquinas SVM para classificar as imagens, como por ex. balastro ou não balastro, outro para deteção de aparelhos de mudança de via (são dispositivos que permitem a passagem dos comboios de uma linha para outra) e um último para detecção de fixações de carril (um dos elementos com maior dificuldade de detecção).

Marino and Stella [18] desenvolveram o ViSyR – A Vision System for Real – Time Infrastructure Inspection. Constitui-se num sistema automático de inspeção visual da via, decomposto em três blocos:

- Deteção do carril através da análise de componentes principais, aplicada à sequência de imagens recolhidas, conseguindo taxas de acerto de 98,5%;
- Deteção de componentes das fixações do carril, desenvolvido em três fases, uma fase de previsão onde identifica imagens candidatas a conter os padrões a serem detetados (identifica as formas geométricas das fixações), uma segunda fase de redução de ruído contido nas imagens captadas e, por último, uma rede neuronal que suporta a classificação das imagens relativamente à presença ou ausência das fixações nas imagens recolhidas. A deteção de componentes das fixações de carril tem uma taxa de acerto de 99,6% na identificação da presença e 95% na deteção da ausência;
- O terceiro bloco, que constitui o ViSyR, visa detetar defeitos na superfície do carril, nomeadamente o efeito ondulatório que existe na sua superfície e que é um dos grandes tópicos relacionados com a manutenção da via, através da aplicação de uma técnica de tratamento de imagem, da qual se retiram médias e variâncias objeto de classificação por uma máquina vetor suporte, conseguindo 100% de eficácia na deteção dos defeitos em apreço.

Outros autores e publicações têm-se debruçado sobre esta temática de recolha de dados, como sejam SHAH [19], relativamente à automatização da inspeção visual a todos os componentes de via. Contudo apesar de já terem sido testados algoritmos para a deteção

de fixações e defeitos superficiais de carril (enrugamentos), apenas era usado de forma recorrente na deteção de defeitos de bitola[6] (afastamento além das medidas pré fixadas entre os carris), Alippi, Casagrande, Scotti and Piuri [20], desenvolve um trabalho relativamente à medição do perfil da Via, Xishi, Bin and Yinhang [21] relativamente à quebra de carril.

Subtema Tratamento dos Dados

Este subtema também é abordado nas publicações supra referidas, que têm como foco principal a recolha de dados. Contudo, existindo publicações que acentuam uma perspetiva face à outra, também optei por efetuar esta destrinça. Desta forma, identifico algumas publicações que utilizam técnicas de *machine learning*, bem como técnicas estatísticas para tratarem problemas relacionados com a gestão do conhecimento/apoio à decisão na infraestrutura ferroviária - Via.

Guclu, Yılboga, Eker, Camci and Jennions [22] apresentam um modelo sobre previsão de falhas em aparelhos de mudança de via com base em modelos auto regressivos de médias móveis. Identificam como escassa a literatura relativamente a modelos de previsão relacionados com falhas em aparelhos de mudança de via, contrariamente a modelos de diagnóstico que caraterizam como um problema de classificação, objeto de várias publicações, como sejam García Márquez, Schmid and Conde Collado [23], Garcia Marquez, Pedregal Tercero and Schmid [24], Roberts, Dassanayake, Lehrasab and Goodman [25], García Márquez and Schmid [26], Atamuradov, Camci, Baskan and Sevkli [27],Márquez, Roberts and Tobias [28].

Camci and Chinnam [29] e Camci and Chinnam [30] esquematizam o problema dos métodos de previsão em três grupos, evolutivos, de estado e de degradação dos prognósticos resultantes da previsão.

[6] Conceito definido no anexo XVIII

Eker, Camci and Kumar [31] publicam um método de diagnóstico para um problema específico dos aparelhos de mudança de via com base em máquinas suporte vetor com *kernel* gaussiano para efeitos de classificação. Definem seis variáveis de análise do comportamento do aparelho de mudança de via que são medidas por um sensor: média, desvio padrão, variância, inclinação, máximo e mínimo. Seguidamente, efetuam uma seleção das variáveis utilizando dois métodos. Primeiro hierarquizando as variáveis através de um teste T, que permite avaliar a eficácia de cada variável na discriminação das classes, escolhendo, subsequentemente as melhores, ou efetuando uma redução de variáveis através de uma análise de componentes principais (ACP). Seguidamente, aplicam uma SVM, quer às variáveis selecionadas através do teste T, quer às selecionadas pela ACP, tendo-se revelado que, após uma amostragem da base de dados através do método *Holdout*, os resultados devolvidos pelo modelo SVM indicaram um maior acerto com as variáveis resultantes da ACP.

Yilboga, Eker, Güçlü and Camci [32], publicam um modelo de predição de falha para o problema identificado no *paper* imediatamente anterior, utilizam uma rede neuronal que, através da adequação da sua estrutura, permite incorporar o fator tempo, e que, alimentada com os dados da degradação no tempo dos aparelhos de mudança de via, depois de convertidos numa escala discreta, produz uma taxa de acerto de 99,3%.

Com este conjunto de exemplos, objeto das publicações acima resumidas, considero que ficou documentado o conhecimento atual sobre as diversas experiências de *Data Mining* ocorridas sobre o contexto da gestão de infraestruturas, nomeadamente ferroviárias.

2.2.3. Hardware/Software Disponível

O Hardware disponível é um computador portátil com 8 GB de RAM e 2.4 Ghz de velocidade de processamento. Este é suficiente para a execução do projeto.

O Software disponível e a utilizar será o SPSS para a análise univariada e bivariada dos atributos, o R para a exploração das bases de dados, o pré-processamento, a implementação e avaliação do algoritmo, o SPSS na análise estatística, o Excel como repositório de dados. Para o desenho de processos será utilizado o software Bizagi. A aplicação CSV SPLIT foi utilizada para decompor os ficheiros de excel acima de 1 milhão de observações.

2.2.4. Requisitos/ Pressupostos/ Constrangimentos/ Riscos

Requisitos

Os requisitos do projeto residem na escolha de um método amigável do ponto de vista da compreensão dos técnicos ligados à manutenção da infraestrutura de forma a que participem na construção dos resultados obtidos, potenciando a aplicação do próprio método na produção de regras interessantes e na utilização como ferramenta organizacional.

Na continuidade da existência das bases de dados trabalhadas em sede de projeto.

No interesse dos especialistas de domínio pelos resultados obtidos pela metodologia.

Pressupostos

O incremento de dados e de base de dados na empresa tenderá a aumentar, nomeadamente na área de manutenção de infraestruturas por via da automatização do processo inspetivo dos diversos componentes de infraestrutura, o que irá alavancar a utilidade do método aqui definido e a introdução de ferramentas de *Data Mining* no geral.

Constrangimentos

Não existir um repositório de informação com o desenho das bases de dados.

Não existir uma normalização de alguns dos repositórios de informação.

Não existir em suporte informático um repositório com todos os elementos de via e seus componentes que permita contextualizar os valores relativos aos parâmetros de via resultantes das inspeções.

Riscos

Inexistência de informação necessária em suporte informático à identificação de padrões interessantes, nomeadamente no que se refere à caraterização do contexto em que a superestrutura de via se insere. Altimetria, planimetria, condições meteorológicas, os próprios elementos que constituem essa superestrutura de via e a caracterização da infraestrutura de via.

O autor da tese não ter conhecimentos técnicos na área.

2.2.5. Terminologia

Através do Glossário elaborado constante do anexo XVIII definem-se os principais conceitos associados ao domínio objeto de estudo.

2.3. Objetivos de *Data Mining*

- Encontrar conjuntos de condições relativos a características da infraestrutura e da sua utilização que estejam associados a distribuições anómalas de valores de indicadores de qualidade da via.
- Encontrar pelo menos um padrão considerado interessante pelo utilizador.
- Aplicar uma metodologia para aferir o interesse da informação devolvida pelo algoritmo de forma a concentrar os especialistas de domínio, que serão os clientes do processo, nos aspetos qualitativos da informação produzida.

3. Conhecimento dos Dados

3.1. Obtenção dos Dados

3.1.1. Seleção das Bases de Dados e Fontes de Informação

Seleção das Bases de Dados Referentes aos Parâmetros Geométricos de Via

Todas as bases de dados referidas no Apêndice I relacionadas com o processo de monitorização dos parâmetros de via são parte de um processo que visa classificar as leituras constantes do ficheiro G_Medições em defeitos e falhas, com o objetivo de intervir na infraestrutura após a deteção da falha de forma a repor os parâmetros geométricos definidos.

Porque o que se pretende é uma abordagem que vise extrair conhecimento das base de dados com vista à possibilidade de se inferir as causas do processo de degradação, desprezaram-se todas as bases de dados que tinham como finalidade obter a classificação de defeitos e falhas tendo-se selecionado o ficheiro G_Medições.

Seleção do Ficheiro Circulação

No ficheiro G_Medições constam apenas o resultado das medições aos parâmetros de via, pretendendo-se identificar hipóteses que relacionem a evolução dos parâmetros da estrutura de via com o contexto envolvente, foi necessário identificar outras fontes de dados que permitissem inferir relações de associação do tipo:

$$\text{Se Componente A} = w \wedge 0{,}5 < \text{Componente B} < 1 \Rightarrow \text{W com defeito}$$

Expressão (3.1): Relação antecedente/consequente

A expressão acima exemplifica uma relação de consequência entre componentes da via A e B, como por exemplo, tipo de travessa, tipo de carril, tipo de fixações, ou outros fatores de contexto, como sejam meteorológicos, geológicos e valores de medições de parâmetros de via W, considerados anómalos.

Esta relação só é possível quando existam bases de dados que relacionem as leituras dos parâmetros de via com o contexto em que estão inseridos. Este pressuposto não se verificou inicialmente dado a inexistência de uma base de dados que os relacionasse.

Desta forma a relação acima (expressão (3.1)) teve que ser construída socorrendo-nos da base de dados circulação denominada T_Tables mencionada no Apêndice I e constante do Anexo I.

Seleção do Ficheiro Diagramas de Via

Como se pode constatar pela leitura do Anexo I, a base de dados circulação contém predominantemente informação relativa ao processo de controlo da circulação. Após consulta aos denominados especialistas do domínio, seria necessário obter mais variáveis pois apenas as existentes não permitiriam efetuar relações de antecedência, consequência, que permitisse efetuar inferências sobre o processo de degradação da infraestrutura.

Não existindo bases de dados disponíveis, existiam contudo suportes com informação relativamente às componentes da infraestrutura suscetíveis de serem constituídos em base de dados, estes denominados de Diagramas de Via, continham informação por ponto quilométrico relativamente a diversos elementos caraterizadores da infraestrutura (ver anexo I).

3.1.2. Seleção da Abrangência Geográfica

Face à limitação provocada pela inexistência de suportes de informação informatizados, no que concerne às componentes da infraestrutura representadas pelos denominados Diagramas de Via, foi necessário circunscrever o âmbito do objeto de estudo a cinco troços, escolhidos em conjunto com os especialistas do domínio.

Troço	Tipo	Quilómetros	Nº de Observações Iniciais
A	Principal	14,875	53 684
B	Principal	3,715	14 861
C	Principal	10,671	42 685
D	Principal	4,577	15 291
E	Secundário	31,697	123 483
Total		**65,535**	**250 004**

Tabela 6: Troços por quilómetros e observações

Generalização e Superajustamento dos Dados

A escolha destes troços procurou que as hipóteses a ser inferidas tivessem capacidade de generalização para que não ocorressem efeitos de superajustamento, ou seja que os troços escolhidos fossem representativos do universo em questão.

3.1.3. Seleção do Período de Observação dos Dados

Critério e Seleção do Período Para os Dados Relativos ao ficheiro G_Medições.

Conforme Andrade [7] vários estudos experimentais validaram uma relação linear entre o desvio padrão do nivelamento longitudinal (filtrado num comprimento de onda de 200 metros), como atributo indicativo do processo de degradação, e o acumulado em toneladas (MGT) suportado pela estrutura, resultante do peso de todos os veículos ferroviários que passaram por determinado troço desde a última ação de renovação ou manutenção.

Segundo o mesmo autor Andrade [7] esta relação pode ser expressa pela seguinte relação linear:

$$\sigma_{LL} = \alpha + \beta.T$$

Expressão (3.2): Regressão linear desvio padrão niv. long. e processo de degradação

Em que σ_{LL} é o desvio padrão dos defeitos longitudinais (mm); α é o desvio padrão inicial após uma renovação ou operação de manutenção (compactação do balastro vulgo "ataque à via"); β é a taxa de deterioração (mm/100 MGT); T é o valor acumulado em MGT desde a última renovação ou operação de manutenção.

Contudo não é possível assumir o critério MGT como elemento definidor do período de observação dos dados a selecionar, pois a recolha de dados (inspeções) é definida por períodos temporais e não atende a este critério.

Poder-se-ia recorrer às leituras pós processo de intervenções na via[7] obtendo o valor de MGT desde a intervenção até à última leitura de inspeção, contudo estas intervenções são diversas e executadas em momentos diferentes ao longo de cada troço, bem como esta informação só recentemente se encontra centralizada não sendo possível em tempo útil a sua identificação para os troços em questão.

O mesmo autor Andrade [7] refere que existiram outras abordagens, igualmente válidas no seu contributo para o processo de estimação da degradação dos parâmetros de via, assentes em outros atributos, velocidade das composições ferroviárias, tempo desde a última intervenção, estrutura da via, e até estudos que concluem que os atributos mais decisivos no processo explicativo da degradação não são identificáveis.

Desta forma face ao exposto, não temos nenhum fator ou atributo que se possa constituir inequivocamente como proxi do processo de degradação e como tal definidor do período de observação dos dados.

Face ao contexto, da ausência de um processo de medição (inspeção) que garanta a fiabilidade de qualquer critério, optou-se por uma simplificação que poderá ser grosseira se assumir o critério MGT como o fator preponderante no processo de degradação de via, mas já se torna mais aceitável se assumirmos que existem outros fatores, tal como

[7] São leituras complementares às semestrais, para assegurar a conformidade dos parâmetros geométricos de via pós intervenções.

supra mencionado, não identificáveis, que poderão ser os principais agentes do processo de degradação da via.

Critério de Seleção

A simplificação reside no garantir um afastamento temporal mínimo relativamente homogéneo para todos os troços entre a data dessas intervenções e a recolha da informação das variáveis supra mencionadas com vista a selecionar o período de observação dos dados relativos aos parâmetros geométricos de via, constantes do ficheiro G_Medições.

Contudo este é um aspeto a melhorar e que deve ser tido em conta quer no momento de leitura dos resultados do algoritmo, quer em futuros desenvolvimentos associados à estruturação das bases de dados subjacentes a esta temática. Esta abordagem deverá ser sistematizada e objeto de operacionalização de forma a se poder selecionar os registos dos parâmetros de via de acordo com a data das intervenções realizadas ou determinado valor de MGT, possibilitando o trabalho da base de dados em função destes dois atributos, afastamento temporal relativamente às intervenções tendentes à correção dos parâmetros de estrutura de via e, ou, de MGT.

Na Figura 2 regista-se o período de observação dos dados relativos aos parâmetros geométricos de via constantes do ficheiro G_Medições, após se ter obtido informações que não existiram intervenções de correção aos parâmetros geométricos de via no primeiro semestre de 2012 nos troços em apreço.

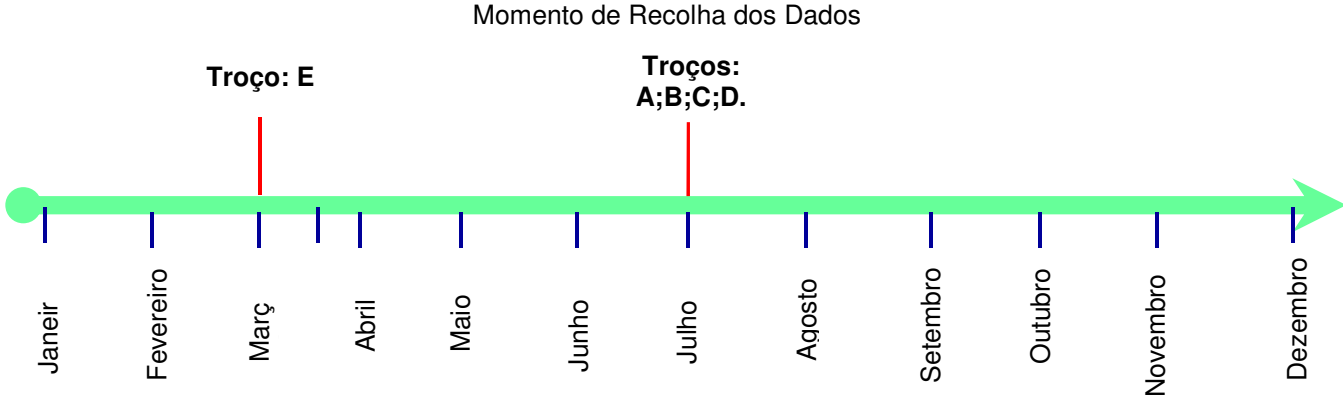


Figura 2: Momento da recolha dos dados

Critério e Seleção do Período Para os Dados Relativos à Circulação.

Definido o período para a seleção dos dados constantes do ficheiro G_Medições é necessário definir o período para os dados constantes do ficheiro Circulação, que período escolher? Os meses que distam desde o início do ano? No caso do troço "E" seriam os 3 primeiros meses, enquanto nos outros troços seriam os primeiros 6 meses, não tendo registo da última medição (inspeção) como garantir que este período temporal expressa uma relação correta entre as variáveis afetas ao ficheiro "Circulação" e as variáveis afetas ao ficheiro G_Medições?

Após consultados os especialistas de domínio assumiu-se que as variáveis contidas no ficheiro Circulação, nomeadamente o atributo MGT (Peso Total) não sofreriam oscilações entre os diferentes períodos anuais, pelo que se poderia assumir o pressuposto que os atributos do ficheiro Circulação seriam atributos caraterizadores dos diferentes troços independentemente da relação temporal com as variáveis afetas ao ficheiro G_Medições, pelo que o espaço temporal escolhido para os dados do ficheiro Circulação foi o ano de 2012.

Conforme já referido esta opção é objeto de crítica e resulta da ausência de uma ferramenta que possibilite retirar em tempo útil os dados relativos aos parâmetros geométricos de via atendendo ao tempo percorrido desde a última intervenção e ou valor de MGT.

3.1.4. Seleção dos Ficheiros

Após definição do período temporal e geográfico a observar, de acordo com o ponto seleção de base de dados, foram selecionados os ficheiros G_Medições, Circulação e criada a Base de Dados Diagrama de Via, conforme apêndice II:

Em conclusão a este ponto foram selecionados os ficheiros com os atributos constantes do anexo II, dando origem ao seguinte número de ficheiros:

	Nº de Ficheiros			
Troço	**Ficheiro G_Medições**	**Ficheiro Circulação**	**Ficheiro Diagramas de Via**	**Total**
Troço E	1	1	1	3
Troço A	1	1	1	3
Troço C	1	1	1	3
Troço B	1	1	1	3
Troço D	1	1	1	3
Total	**5**	**5**	**5**	**15**

Tabela 7: Nº de ficheiros selecionados

3.2. Descrição dos Dados

Esta secção respeita à descrição e caraterização dos dados presentes em cada tipo de ficheiros selecionado na subsecção anterior.

3.2.1. Caraterização dos Atributos

Conforme podemos constatar nos quadro infra os atributos dos ficheiros estão classificados relativamente a cinco caraterísticas[8], que visam caraterizá-los relativamente a:

a. Facilidade de obtenção dos atributos, medido em tempo necessário e ou complexidade.
b. Tipo de atributo.
c. Relevância para o objetivo de *Data Mining*, obtenção de padrões interessantes.

[8] Caraterísticas definidas na metodologia CRISP-DM [4] Chapman, J. C. R. K. T. K. T. R. C. S. R. W. P. *CRISP-DM 1.0 - Step-by-step data mining guide*. SPSS, 2000.

d. Consenso/entendimento do seu significado na empresa.
e. Relevância segundo o especialista de domínio.

Relativamente às caraterísticas "a." e "e." foi utilizada uma escala de Likert[9], com rótulo, do tipo unipolar, com número ímpar de opções e inclusiva.

Acessibilidade dos Atributos	
Inacessiveis	0
acessiveis com dificuldade	1
acessiveis	2
acessiveis com facilidade	3
Imediato	4

Tabela 8: Escala de acessibilidade dos atributos

Relevância dos Atributos Especialista	
Irrelevante	0
Pouco Relevante	1
Relevante	2
Muito Relevante	3
Determinante	4

Tabela 9: Escala de relevância dos atributos

Ficheiro G_Medições

No ficheiro G_Medições cada observação corresponde um ponto da estrutura ferroviária de via (Via), e resulta das leituras da inspeção automatizada que, de 25 cm em 25 cm, obtém um conjunto de parâmetros cujos valores não são absolutos, mas valores diferenciais face a valores referenciais padrão pré definidos, que representam a fiabilidade da infraestrutura. Este ficheiro é composto pelas varáveis infra expostas e com o significado constante do anexo III.

[9] Escala de Likert com rótulo significa que cada valor é associado a uma expressão, unipolar significa que a extremidade de cada escala é o oposto da outra, inclusiva significa que abrange todo o conjunto de respostas possíveis.

G_Medições (Inicial)	Acessibilidade dos Atributos	Tipo de Atributo	Atributo Relevante para Obj. Data Mining	Atributo com Significado Constante (na empresa)	Relevância Segundo o Especialista de Domínio
ANO	2	Numérico	N	S	0
ID_INSP	2	Numérico	N	S	0
DIST_O	2	Numérico	N	S	0
ID_POS	2	Numérico	N	S	0
KM	2	Numérico	N	S	0
LOCALIZ	2	Numérico	N	S	0
PK	2	Numérico	N	S	0
P_SPEED	2	Numérico	N	S	0
F_SPEED	2	Numérico	N	S	0
SPEED	2	Numérico	N	S	0
NIVLE	2	Numérico	S	S	4
NIVLD	2	Numérico	S	S	4
NIVLED1	2	Numérico	S	S	4
NIVLDD1	2	Numérico	S	S	4
NIVLED2	2	Numérico	S	S	1
NIVLDD2	2	Numérico	S	S	1
ALINE	2	Numérico	S	S	4
ALIND	2	Numérico	S	S	4
ALINED1	2	Numérico	S	S	4
ALINDD1	2	Numérico	S	S	4
ALINESQR	2	Numérico	S	S	3
ALINDIRR	2	Numérico	S	S	3
NIVTRANS	2	Numérico	S	S	2
NIVTRANR	2	Numérico	S	S	2
EMPENO3M	2	Numérico	S	S	4
EMPREL3M	2	Numérico	S	S	4
EMPENO9M	2	Numérico	S	S	3
EMPREL9M	2	Numérico	S	S	3
BITOLA	2	Numérico	S	S	4
BITMED	2	Numérico	S	S	4
GRAD	2	Numérico	S	S	3
GRADMED	2	Numérico	S	S	3
CURVA	2	Numérico	S	S	3

Tabela 10: Classificação dos atributos ficheiro G_Medições

Ficheiro Circulação

O ficheiro Circulação expressa todas as circulações ocorridas e respetivos atributos caraterizadores num determinado espaço geográfico (linha/troço) entre duas datas à escolha. Este ficheiro é composto pelas varáveis infra expostas

Circulação (Inicial)	Acessibilidade dos Atributos	Tipo de Atributo	Atributo Relevante para Obj. Data Mining	Atributo com Significado Constante (na empresa)	Relevância Segundo o Especialista de Domínio
NumeroComboio1	3	Numérico	N	S	0
NumeroComboio2	3	Numérico	N	S	0
DataRealizacao	3	Numérico	N	S	0
OperadorId	3	Numérico	N	S	0
Operador	3	Categórico	N	S	0
RegimeFrequenciaMnemonica	3	Categórico	N	S	0
TipoServicoId	3	Numérico	N	S	0
TipoServico	3	Categórico	N	S	0
DataInicioCirculacao	3	Numérico	N	S	0
DataFimCirculacao	3	Numérico	N	S	0
DocumentoHorario	3	Categórico	N	S	0
DependenciaOrigemId	3	Numérico	N	S	0
DependenciaOrigemDescricao	3	Categórico	N	S	0
DependenciaDestinoId	3	Numérico	N	S	0
DependenciaDestinoDescricao	3	Categórico	N	S	0
HoraPartida	3	Numérico	N	S	0
HoraChegada	3	Numérico	N	S	0
IsSuprimidoTotal	3	Categórico	N	S	0
IsSuprimidoParcial	3	Categórico	N	S	0
CargaRebocada	3	Numérico	S	S	3
ComprimentoRebocado	3	Numérico	S	S	2
IdMaterialMotor1	3	Numérico	S	S	2
VelocidadeMaxima	3	Numérico	S	S	2
Sentido	3	Categórico	S	S	0

Tabela 11: Classificação dos atributos ficheiro Circulação

Ficheiro Diagrama de Via

O ficheiro Diagrama de Via tal como referido no ponto seleção dos ficheiros teve origem na transposição para suporte informático de dados que se encontravam em suporte físico. Esta fonte de informação documenta a estrutura de via em altimetria e planimetria fazendo referência a alguns componentes da superestrutura e infraestrutura.

Após consulta aos especialistas de domínio, foram identificados e transpostos para suporte informático os atributos infra expostos, que foram indicados como aqueles com melhor relação entre esforço despendido na sua transformação para suporte informático

e resultados obtidos, no que concerne a caraterizar a evolução dos diversos parâmetros geométricos de via em estudo.

Desta forma, dos atributos constantes do anexo I, referente aos "Diagramas de Via" foram selecionados os seguintes atributos:

- Inclinação;
- Carril;
- Carril II;
- Travessa;
- Velocidade Máxima de Planta.

À Semelhança dos outros ficheiros estes atributos são classificados da seguinte forma:

Diagrama de Via	Acessibilidade dos Atributos	Tipo de Atributo	Atributo Relevante para Obj. Data Mining	Atributo com Significado Constante (na empresa)	Relevância Segundo o Epecialista de Domínio
PK	1	Numérico	N	S	0
Inclinação	1	Numérico	S	S	2
Curva de Concordância	0	Numérico	S	S	2
Carril	2	Numérico	S	S	4
Carril II	2	Categórico	S	S	4
Travessa	2	Categórico	S	S	4
Vel. Máxima de Planta	2	Numérico	S	S	3

Tabela 12: Classificação dos atributos ficheiro Diagramas de Via

3.2.2.Volumetria dos Ficheiros

G_Medições

G_Medições		
Troço	**Nº Observações**	**Nº Atributos**
Troço A	53 684	33
Troço B	14 861	33
Troço C	42 685	33
Troço D	15 291	33
Troço E	123 483	33
Total	**250 004**	**33**

Tabela 13: Volumetria G_Medições

Circulação

Circulação		
Troço	**Nº Observações**	**Nº Atributos**
Troço A	10 048	24
Troço B	26 305	24
Troço C	28 037	24
Troço D	30 748	24
Troço E	3 871	24
Total	**99 009**	**24**

Tabela 14: Volumetria Circulação

Diagramas de Via

Diagramas de Via		
Troço	**Nº Observações**	**Nº Atributos**
Troço A	53 684	7
Troço B	14 861	7
Troço C	42 685	7
Troço D	15 291	7
Troço E	123 483	7
Total	**250 004**	**7**

Tabela 15: Volumetria Diagramas de Via

3.2.3. Decisão de Integração

Face à volumetria dos ficheiros identificada na subsecção anterior, associada à existência de processos manuais na inserção de observações (Diagramas de Via), optou-se por uma integração dos ficheiros em etapas, numa abordagem troço a troço, pois permite uma maior capacidade de deteção de erros, uma vez que a sua dimensão é não só menor como também circunscrita a um espaço com características mais homogéneas. Esta abordagem também permite uma maior flexibilidade na eventual reconstrução do ficheiro final ou exploração dos dados por troço.

3.2.4 Definição da Estratégia Relativa à Exploração e Preparação dos Dados

Face aos atributos do ficheiro circulação não serem diretamente integráveis[10], foi necessário efetuar para estes ficheiros e para cada um dos troços, a exploração e preparação dos dados antes do processo de integração.

Relativamente aos ficheiros G_Medições e Diagramas de Via, não existe à partida interesse no conhecimento da distribuição associada de cada um dos atributos para cada um dos troços mas sim para as distribuições atendendo ao universo dos cinco troços em estudo.

Desta forma, atendendo ao supra mencionado, a análise exploratória dos dados e preparação dos ficheiros (conforme subpontos 2 e 3 da metodologia CRISP-DM Chapman [4] será elaborada em dois momentos, primeiro para o ficheiro Circulação e depois para o ficheiro integrado da seguinte forma:

[10] Existe uma relação com os outros ficheiros de muitos para muitos, para melhor entendimento ver ponto transformação de atributos.

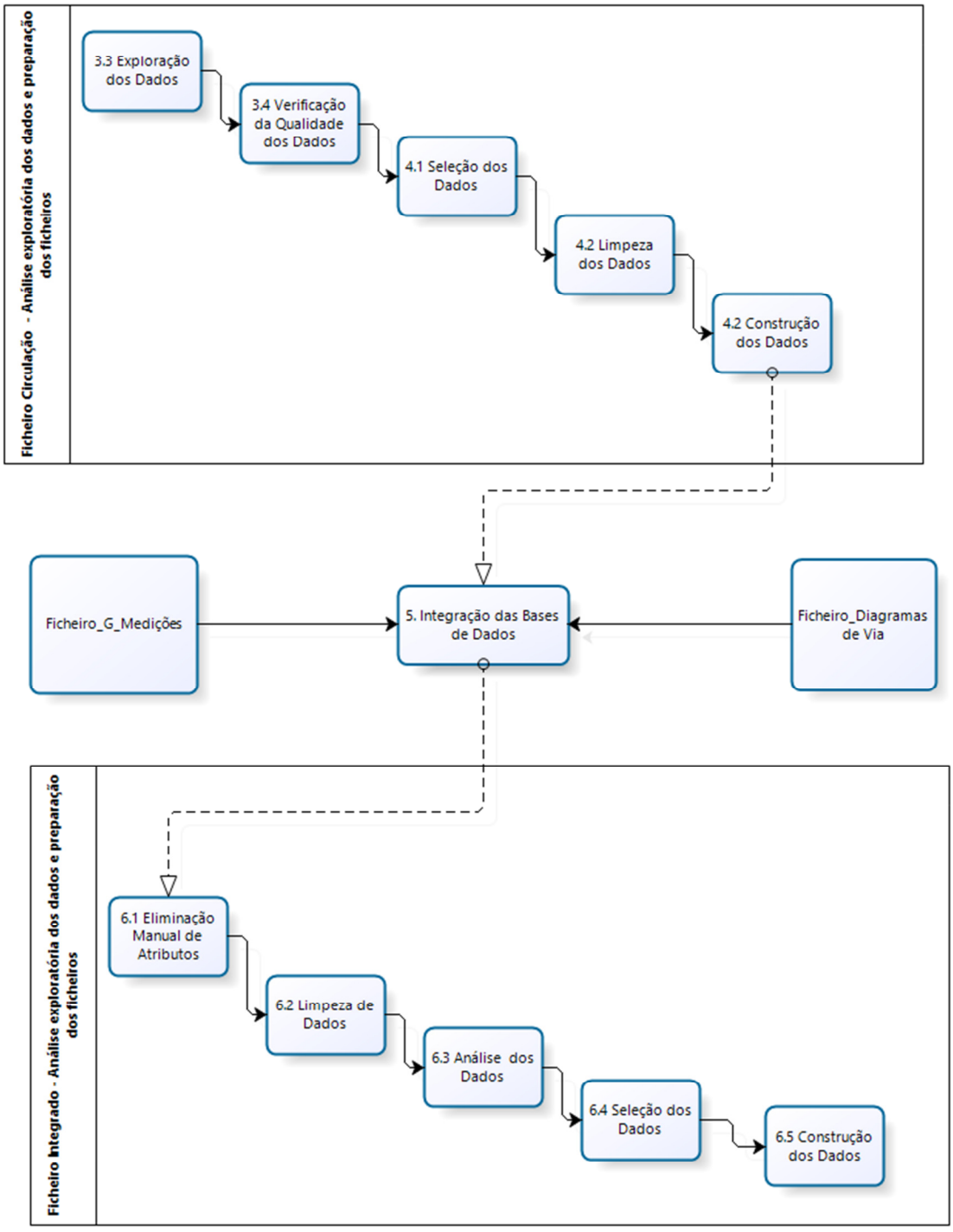


Figura 3: *Workflow* conhecimento e preparação dos dados

3.3. Análise dos Dados - Ficheiros Circulação

Nesta fase a análise dos dados prendeu-se com a identificação através de inspeção visual dos atributos relevantes identificados na tabela 11, tendo-se selecionados os seguintes atributos:

- Carga Rebocada;
- Comprimento Rebocado;
- Velocidade Máxima.

3.4. Qualidade dos Dados - Ficheiros Circulação

A qualidade dos dados neste ficheiro foi verificada da seguinte forma:

1. Garantindo que a data das observações dizem respeito à data pretendida, e as linhas/troços de linha são também os selecionados. Efetuado através de um simples filtro em Excel para o atributo data garantindo a coerência da data com o período definido:

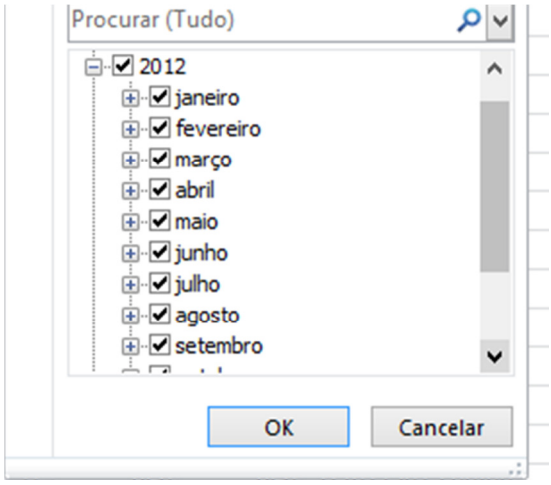


Figura 4: Filtro (Excel) datas

2. Garantindo que nos atributos escolhidos não existe ruído:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CargaRebocada	3873	-1,00	970,00	22,7382	112,88635
ComprimentoRebocado	3874	-1,00	370,00	12,7126	47,58636
VelocidadeMaxima	3874	-1,00	120,00	113,3730	19,18389
Valid N (listwise)	3873				

Tabela 16: Estatísticas descritivas - Circulação

Como se pode verificar existem valores negativos para todos os atributos, pelo que atendendo à natureza de cada atributo, são classificados como ruído. Os limites máximos foram verificados à luz da especialidade de domínio sendo aceites. O ruído foi tratado no ponto 4.1.

3. Garantindo que nos atributos escolhidos não existem casos omissos:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
CargaRebocada	3873	91,0%	383	9,0%	4256	100,0%
ComprimentoRebocado	3873	91,0%	383	9,0%	4256	100,0%
VelocidadeMaxima	3873	91,0%	383	9,0%	4256	100,0%

Tabela 17: Validação de casos omissos - Circulação

Como se pode verificar os casos omissos representam no caso do ficheiro em apreço 9%, estes também serão objeto de tratamento em sede de seleção de dados.

4. Preparação dos Dados – Ficheiro Circulação

4.1. Seleção dos Dados

Observado que os campos carga e comprimento, previamente selecionados em sede de exploração dos dados, apenas diziam respeito às composições, não agregando os valores da automotora, foi necessário juntar ao ficheiro Circulação os atributos Peso Bruto e Comprimento referentes à automotora, provenientes de um ficheiro denominado T_Material_Motor , através do atributo chave Id_Material Motor1.

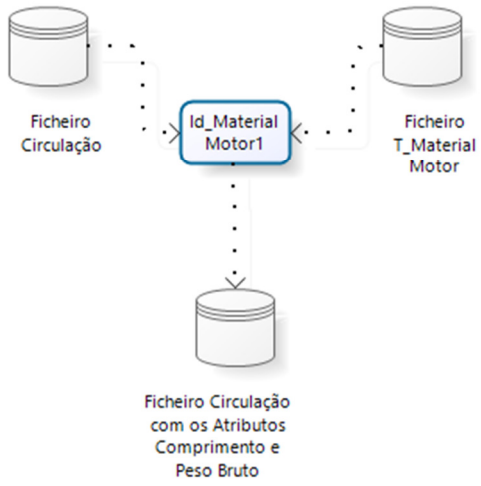


Figura 5: Aquisição de atributos - Circulação

Desta forma e atendendo ao "novo" ficheiro Circulação foram novamente selecionados os seguintes atributos:

- Carga Rebocada;
- Comprimento Rebocado;
- Comprimento;
- Peso Bruto;
- Velocidade Máxima.

4.2. Limpeza dos Dados

Dados com Ruído

Relativamente aos dados com ruído identificados na secção 3.4 “Qualidade dos dados”, decidiu-se efetuar a sua supressão em sede de transformação de variáveis pois identificou-se que tinham origem em observações que registaram só a passagem da automotora atribuindo aos campos o valor de “-1”, logo quando os valores referentes à mesma foram adicionados deixou de existir ruído.

Casos Omissos

Relativamente às observações com valores omissos, apenas foram identificadas no ficheiro relativo ao Troço E, tendo-se optado pela eliminação das observações.

Após a limpeza dos dados em todos os ficheiros Circulação ficámos com o seguinte nº de observações:

Circulação		
Troço	**Nº Observações Iniciais**	**Nº Observações**
Troço A	10 048	10 048
Troço B	26 305	26 305
Troço C	28 037	28 037
Troço D	30 748	30 748
Troço E	4 257	3 871
Total	**99 395**	**99 009**

Tabela 18: Ficheiro Circulação após limpeza dos dados

4.3. Construção dos Dados

Transformação dos Atributos

Dados Redundantes

Os atributos selecionados no ponto 4.1 têm que ser combinados por forma a se obter um novo atributo, pois apresentam a mesma informação preditiva.

Desta forma os atributos "Carga Rebocada" e "Peso Bruto", foram somados em cada observação dando origem a um novo atributo denominado de "Carga_Total".

$$Carga\ Total = \sum_{1}^{n} (Carga\ Rebocada + Peso\ Bruto)$$

Expressão (4.1): Cálculo do atributo "Carga Total"

Idêntico processo ocorreu com os atributos "Comprimento Rebocado" e "Comprimento", dando origem a um novo atributo denominado "Comprimento Total".

$$Comprimento\ Total = \sum_{1}^{n} (Comprimento\ Rebocado + Comprimento)$$

Expressão (4.2): Cálculo do atributo "Comprimento Total"

Cumpre referir que os problemas de ruído identificados na secção "Qualidade dos Dados" foram suprimidos com a operação da soma supra exposta, através de uma operação condicional aplicada a cada uma das observações que garante um valor positivo pois os valores provenientes do ficheiro T_Material_Motor não apresentavam ruído.

Dados Inconsistentes

Os valores constantes destes atributos, "Carga Total"; "Comprimento Total"; "Velocidade Máxima", estavam expressos relativamente ao ficheiro G_Medições"[11]

[11] Explicação complementada no ponto Integração das Bases de Dados

numa relação de muitos para muitos , encontravam-se por observação de circulação (“comboio”), necessitando de obterem algum tratamento estatístico para podermos obter o valor representativo de cada atributo para o troço e período temporal em causa.

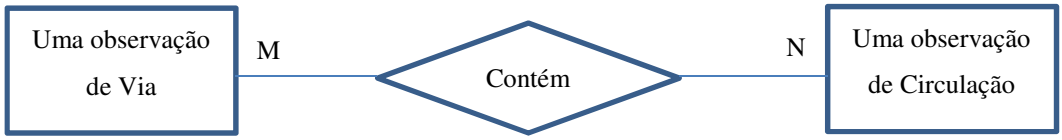


Figura 6: Relação entre observações G_Medições / Circulação

Desta forma foi efetuado um processo de exploração de dados por forma a identificar-se quais os valores de localização ou de tendência central a constar em cada um destes atributos, para cada troço/observação de via, e para o período temporal em apreço.

Carga Total

De acordo com os especialistas de domínio, o peso total que a estrutura suportou é uma referência importante, pelo que foi somado para cada troço a “Carga Total” de cada observação (“comboio”) e atribuído esse valor ao atributo “Peso Total” que carateriza cada um dos troços a integrar no ficheiro Matriz Final.

$$Peso\ Total = \sum_{1}^{n} Carga\ Total$$

Expressão (4.3): Cálculo do atributo “Peso Total”

Relativamente aos atributos “Comprimento Total” e “Velocidade Máxima” pretendendo obter um valor de localização ou de tendência central que caraterizasse cada um dos atributos em cada troço, recorri à estatística descritiva. Conforme Apêndice III selecionaram-se os seguintes valores:

Comprimento Total

Para este atributo foi escolhida a mediana, que coincide com o valor modal para as séries em apreço conforme Tabela 19, infra, como valor de referência para caraterizar o atributo "Comprimento Total".

Comprimento	Mediana	Valor Modal	Frequência
Troço A	66,8	66,8	84%
Troço D	66,8	66,8	60%
Troço E	70	70	86%
Troço C	66,8	66,8	65%
Troço B	66,8	66,8	64%

Tabela 19: Mediana e valor modal atributo "Comprimento Total" por troço

Velocidade Máxima

Para este atributo foi escolhida a média para cada um dos troços conforme Tabela 20

Troço	Média
Troço A	150
Troço D	146
Troço E	113
Troço C	150
Troço B	151

Tabela 20: Média atributo "Velocidade Máxima" por troço

Desta forma e relativamente ao ficheiro "Circulação" conclui-se o processo de conhecimento e preparação dos dados, destinado a escolher as variáveis e respetivos valores a serem incorporados no ficheiro a integrar. Ficheiro Circulação Inicial, Intermédio e Final conforme anexo IV.

5. Integração das Bases de Dados

Concluída para cada um dos troços a seleção dos ficheiros Circulação, Diagramas de Via e G_Medições, conforme anexo V, efetuada a exploração e limpeza dos dados relativos aos ficheiros Circulação, encontramo-nos em posição de efetuar a sua integração.

Integração dos Ficheiros

Com vista a facilitar a integração dos ficheiros constantes do anexo V, estes são integrados troço a troço tendo por base os valores do campo "PKTrabalhado", este referencial é trabalhado de acordo com o processo definido no Apêndice IV, dando origem a um ficheiro único por troço caraterizado pelos atributos representados no anexo VI:

Integração dos Cinco Troços num só Ficheiro

Conforme fig. 13 foram integrados os cinco ficheiros relativos aos cinco troços, nesta etapa apenas se verificou uma colagem das observações dos cinco troços, para um ficheiro comum.

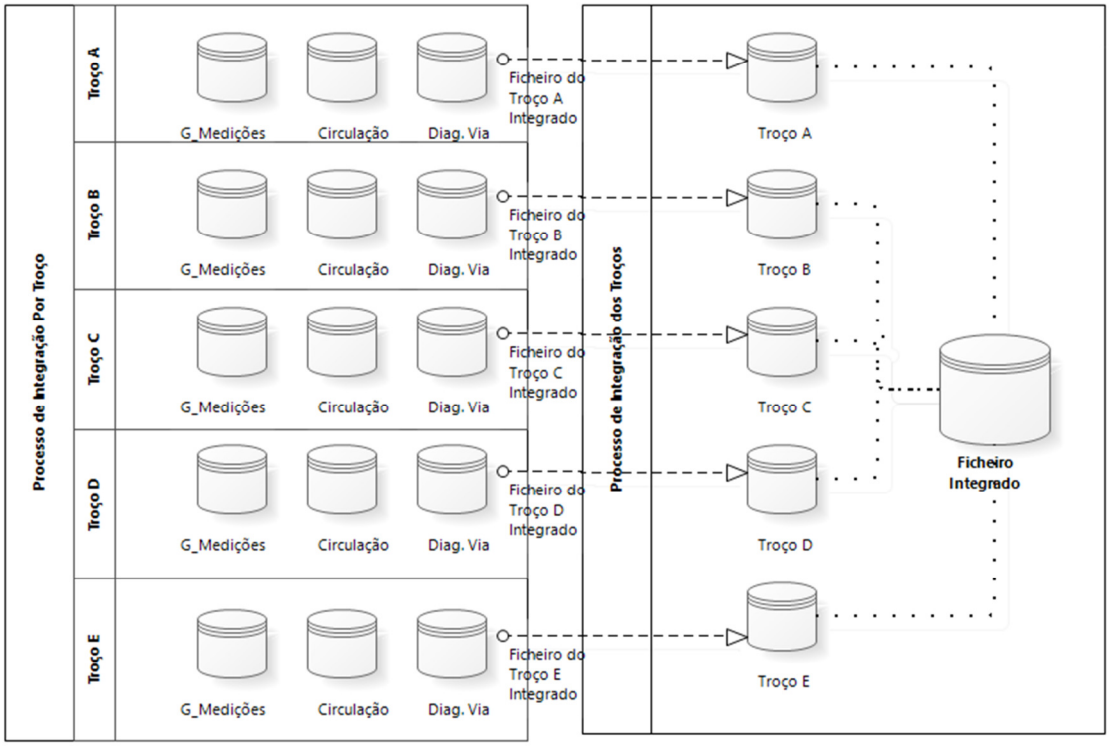


Figura 7: Processo de integração dos ficheiros

Efetuada a integração dos quinze ficheiros relativos aos "Diagramas de Via", G_Medições e "Circulação" correspondentes a cada um dos cinco troços obtemos o Ficheiro Integrado com os atributos constantes do anexo VI, que será objeto de exploração e preparação dos dados.

6. Análise, Qualidade e Preparação dos Dados - Ficheiro Integrado

De forma a não sobrecarregar a análise de dados com atributos irrelevantes efetuou-se uma eliminação prévia dos mesmos antes da referida análise.

6.1. Eliminação Manual de Atributos - Ficheiro Integrado

Com base na caraterização dos atributos efetuada na secção 3.2, de acordo com a tabela 10, foram eliminados do ficheiro agregador dos cinco troços (Ficheiro Integrado), constante do anexo VI, todos os atributos que se consideraram irrelevantes para o *Data Mining* e para os especialistas de domínio, inclusive dois atributos de suporte criados para apoiar o processo de integração não constantes do ficheiro inicial Figura 8.

Atributos Eliminados
ANO
ID_INSP
DIST_O
DIST_O_Trab
ID_POS
KM
LOCALIZ
PK
PKTrabalhado
Confirmação
P_SPEED
F_SPEED
SPEED

Figura 8: Atributos eliminados Ficheiro Integrado

Assim obtemos o ficheiro Integrado Final conforme anexo VI.

6.2. Limpeza de Dados – Ficheiro Integrado

Seguidamente efetuou-se uma análise exploratória dos dados em R com vista a detetar atributos incompletos (sem valores). Conforme anexo VII podemos visualizar que dos

trinta e dois atributos, vinte e quatro apresentavam observações com valores "NA'S" (atributo sem valor). Contudo os atributos incompletos não se ficavam apenas pelos 24 atributos identificados com valores "NA'S" existiam também valores omissos, não identificados com "NA'S", no atributo Inclinação e Curva de Concordância.

A repartição dos valores omissos é desigual entre os atributos, sendo que a maior parte apresentam um valor entre os 600 e os 1.500 valores omissos o atributo Vel. Máxima de Planta assinalava 47.578 valores omissos, e embora não assinalado na figura constante do anexo VII como "NA", tanto o atributo Inclinação como Curva de Concordância assinalam valores omissos de 69.508 e 47.578 respetivamente.

Os 47.578 valores omissos dizem respeito a um desajustamento entre o ficheiro G_Medições e o ficheiro Diagramas de Via do Troço E, nomeadamente a inexistência de atributos no ficheiro Diagrama de Via para as observações do ficheiro G_Medições. Os remanescentes valores omissos dividem-se por 21.930 observações relativos à inexistência de valores para o atributo Curva de Concordância, e 1581 relativos aos atributos parâmetros geométricos de via conforme figura infra:

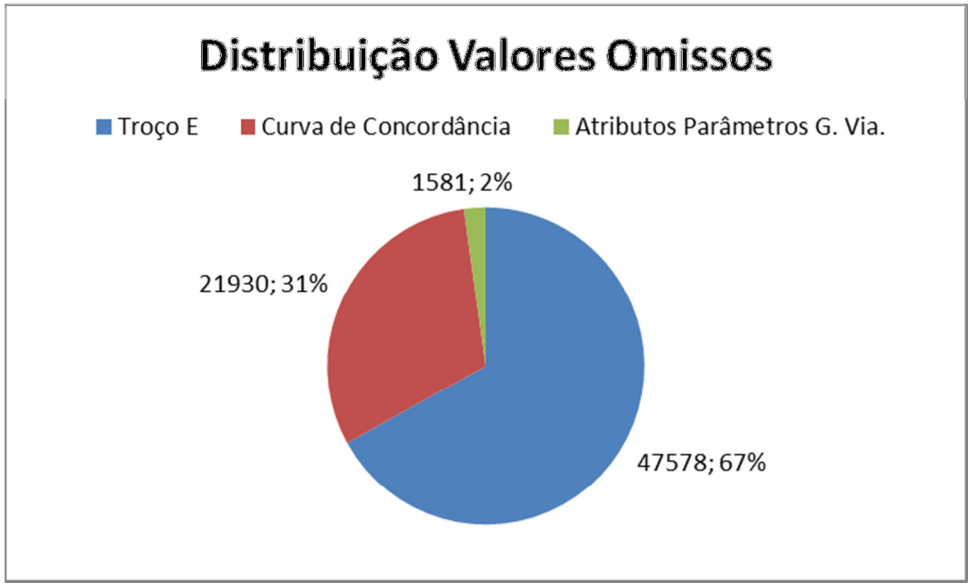


Figura 9: Distribuição de valores omissos

Tendo em conta que 98% dos valores omissos estão circunscritos aos factos apresentados, e sendo os casos que realmente poderiam interferir de forma qualitativa com a estrutura do ficheiro residuais (2%), eliminou-se, recorrendo ao R, todas as observações com valores omissos prosseguindo com o ficheiro com as 178.915 observações conforme quadro infra:

Troço	Nº Observações Iniciais Ficheiro Integrado	Nº Observações Finais (Retiradas observações com atributos omissos)	% Observações Restantes
A	53.684	43.492	81%
B	14.861	14.149	95%
C	42.685	35.040	82%
D	15.291	13.207	86%
E	123.483	73.027	59%
Total	**250004**	**178915**	**72%**

Tabela 21: Observações iniciais e finais por troço – Ficheiro Integrado

Tendo-se retirado todas as observações que não continham qualquer valor para o atributo Inclinação que correspondiam a todas as observações que se encontravam em curva de concordância, foi também eliminado o atributo Curva de Concordância[12] .

Desta forma eliminados todos os atributos irrelevantes à luz dos critérios supramencionados e os casos omissos, o processo de exploração de dados ficou centrado (nos atributos relevantes) e reduzido (à informação com valores).

[12] Curva de concordância são segmentos de linha que ligam dois segmentos com valores de altimetria diferenciados

6.3. Análise dos Dados – Ficheiro Integrado

Esta secção serve para conhecer a base de dados em presença nomeadamente que tipo de dados temos e qual o seu potencial para extração de conhecimento, quais os padrões mais relevantes e eventuais distribuições subjacentes ao processo de geração de dados.

Caraterização do Tipo e Escala dos Atributos

Conforme anexo VIII das 31 variáveis em presença 27 são do tipo numérico e de escala de razão, sendo os remanescentes atributos do tipo qualitativo de escala nominal.

Aparentemente este conjunto de dados é rico em informação, não só por a maior parte dos atributos serem numéricos como, sendo predominantemente de escala de razão, os números possuírem um sentido absoluto. A utilizarmos algoritmos e ou métodos que permitam a interpretabilidade nos resultados[13], possibilita uma maior facilidade de leitura pelo utilizador/especialista de negócio e portanto de validação dos resultados do processo.

Contudo analisados os atributos numéricos bem como os qualitativos no contexto da base de dados verificamos:

Conforme anexo IX os atributos subdividem-se entre 20 atributos relativos aos parâmetros da estrutura de via (consequentes) e 11 relativos às características de via (antecedentes) .

Conforme anexo X dos 11 atributos antecedentes apenas 3 atributos, Grad, Curva e Inclinação poderão conter potencial de informação, que permita pela variabilidade de valores que apresentam, obter diversidade de regras de distribuição (o quarto atributo

[13] Se por exemplo utilizássemos técnicas de redução de variáveis como a análise ACP, em que são criadas novas variáveis substituindo as originais, esta interpretabilidade possibilitada pelos atributos serem da escala de razão, perder-se-ia.

GradeMed apesar de apresentar diversidade de valores é valor médio do atributo Grad e portanto apresenta elevada correlação com este).

Os remanescentes 7 atributos são pouco informativos quando atendemos à diversidade de valores que encerram, 2 qualitativos, normalmente menos informativos que os de escala numérica e 5 que apesar de numéricos e de escala de razão, apresentam reduzida diversidade de valores.

Análise Univariada

Cumpre referir que a análise univariada, com exceção do atributo Inclinação, foi efetuada com base num ficheiro intermédio com 200.698 observações. Esta opção teve como intuito aproveitar o maior número de observações existentes para caracterizar as distribuições associadas, pois a diferença para o ficheiro final de 178.915 observações prende-se exclusivamente com valores inexistentes para o atributo Inclinação.

Uma análise univariada deve atender a um conjunto de medidas que permitam caracterizar a distribuição. Esta caracterização fica facilitada se encontrarmos à priori o tipo de distribuição associada aos dados. Desta forma a análise univariada, separando os atributos por antecedentes e consequentes, atendeu aos pontos infra, mesmo que nem sempre de forma expressa no presente texto.

- Efetuar Teste à Normalidade da Distribuição;
- Medidas de Centralidade;
- Medidas de Dispersão;
- Medidas de Forma.

Variáveis Antecedentes

Tendo-se concluído que dos 11 atributos caracterizadores de via só três apresentam diversidade de valores nas suas observações, só nestes se justifica uma análise univariada à sua distribuição.

Desta forma como podemos visualizar nos histogramas constantes do anexo XI a diversidade de valores que apresentam não é comportável com a sua qualidade de atributos "antecedentes" pelo que terão que ser discretizados. Assim apenas se publicam na parte inferior do anexo XI os valores relativos à estatística descritiva, para concluir que se rejeita a hipótese de normalidade das distribuições. Este facto é confirmado pelos valores de curtose e assimetria, estando em presença de distribuições mais altas e concentradas que as distribuições normais.

Esta análise demonstra que no atributo Inclinação existe uma frequência muito alta de observações no valor zero, relativamente às demais, o que significa que uma parte substancial das observações ocorreram em patamar (num plano não inclinado). O mesmo ocorre no atributo "curva", a concentração desenvolve-se no ponto zero significando que a maior parte das observações foram efetuadas em reta.

Variáveis Consequentes

Efetuada a análise à normalidade da distribuição das variáveis através do teste de Kolmogorov Smirnov (KS), conforme anexo XII, conclui-se que sendo o p.value inferior a qualquer valor de α rejeita-se a hipótese da normalidade de todas as distribuições. Não conhecendo a função densidade probabilidade não conseguimos efetuar o processo estatístico de inferência que nos permitiria assegurar o processo que gera os dados. Desta forma atende-se às principais medidas caracterizadoras das distribuições em presença.

Conforme anexo XIII podemos observar que relativamente à forma das distribuições os valores de assimetria são relativamente próximos de zero, o que significa que as distribuições são aproximadamente simétricas (Skewness). Contudo já relativamente à curtose os valores apresentados indicam uma concentração superior e com maior frequência que uma distribuição normal o que nos traz uma indicação positiva sobre o estado da estrutura de via em apreço, pois tratando-se de atributos diferenciais face ao valor óptimo, valor zero em termos de desvio, sempre que a maior frequência destes

valores se verifica no valor zero significa que a maior parte das observações não revelaram qualquer anomalia do parâmetro em causa.

Se atendermos ao anexo XIV as inferências supra mencionadas verificam-se na observação dos histogramas, semelhantes ao processo de um atributo com distribuição normal só que com maior concentração e mais elevadas. Contudo os atributos assinalados com um círculo, ALINE, ALIND, NIVTRANS, EMPENO 9M, BITOLA e BITMED fogem à distribuição dos demais pela maior dispersão dos quatro primeiros, e assimetria positiva dos dois últimos o que pode indicar um desvirtuamento no controlo destes parâmetros, respetivamente, face aos valores de tolerância pré estabelecidos para o desvio padrão, e face aos valores de referência central. Este valor de referência central deveria assumir o valor zero significando a ausência de desvio, o que se traduz na prática que a variável (BITOLA e BITMED) apresentam um conjunto alargado de observações com um desvio significativo face ao valor zero de referência, que no domínio de conhecimento em estudo denomina-se de excesso de bitola.

Análise Bivariada

Uma questão relevante que se coloca na base de dados em apreço é capacidade de generalização das relações que se venham a estabelecer, nomeadamente em que medida é que as características dos troços, a sua idiossincrasia de contexto, impactam nas relações inferidas. Ou seja, estabelecida uma relação, ela depende exclusivamente dos antecedentes e consequentes em presença ou é fruto destes mas apenas no contexto de determinado troço.

Introdução de Novo Atributo no Ficheiro

Desta forma aproveitando a variável peso total que a cada valor corresponde um e apenas um só troço conforme tabela 22, introduzi um novo atributo no ficheiro, deste atributo consta o nome do troço a que cada observação diz respeito.

Seguidamente efetuei uma análise bivariada, com a distribuição dos atributos quantitativos, correspondentes aos consequentes (parâmetros de via), pelas diferentes classes do atributo qualitativo troço, conforme anexo XV.

Troço	Peso Total*
Troço E	528 482
Troço A	1 541 135
Troço C	6 271 537
Troço B	7 683 468
Troço D	7 778 428

* Toneladas

Tabela 22: Atributo "Peso Total" por troço

Atendendo ao anexo XV podemos observar que existem diferenças entre a distribuição dos atributos ligados aos parâmetros pelos diferentes troços, estas são sobretudo a nível da dispersão dos dados, contudo não se consegue obter um padrão afirmando que determinado troço tem uma maior ou menor dispersão face aos demais.

Relativamente a medidas de localização de tendência central, observando o anexo XV. é possível verificar que a forma das distribuições é aproximadamente simétrica, contudo o peso de outliers que se verificam em todas as distribuições poderá distorcer a média como medida de localização central, o que estando em presença de variáveis quantitativas permite-nos afirmar que sendo a mediana medida de referência de tendência central mais resistente que a média, e sendo esta relativamente equivalente entre todos os troços, implica que a medida de localização central é a mesma nos diferentes atributos para os diferentes troços com exceção dos atributos "BITOLA", BITOLAMED", "GRAD" e "GRADMED".

Os considerandos desta análise bivariada podem ser concretizados no domínio prático atendendo a três variáveis consideradas como as mais relevantes no domínio da expertise em estudo, "EMPENO3M", "NIVTRANS" e "BITOLA", com interrogações de índole operacional como sejam, o que justifica que o Troço A tenha uma menor dispersão dos dados em dois dos três atributos em estudo? Relativamente ao atributo

"BITOLA" o que justifica o excesso de bitola existente no Troço E? Qual a razão deste troço apresentar uma distância interquartis (onde se concentram as observações centrais, 50% do total) sempre superior aos demais nos atributos em apreço? Existem condições estruturais que o justifiquem? Estas observações devem estar presentes no momento de avaliação de resultados do algoritmo, por forma a perceber qual o impacto do contexto nas relações de antecedentes e consequentes estabelecidas.

6.4. Seleção dos Dados – Ficheiro Integrado

Seleção dos Dados Atendendo ao Tipo e Escala de Atributo

Conforme referido na secção 6.3 existem um conjunto de atributos consequentes (associados às caraterísticas/contexto de via) que face à reduzida diversidade de valores, são pouco informativos, podendo desta forma ser candidatos à eliminação. Contudo estes atributos quando combinados entre si apresentam uma distribuição exponencial relativamente ao número de combinações que podem formar em sede de antecedentes, conforme Figura 10, pelo que mesmo os com menor potencial de informação permaneceram na base de dados.

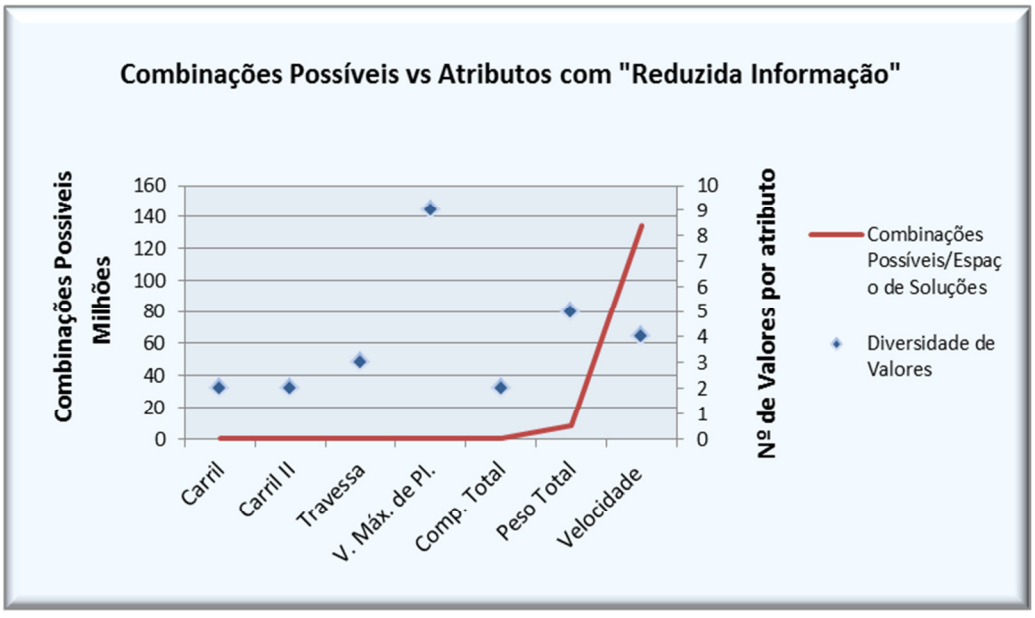


Figura 10: Espaço de soluções vs atributos antecedentes

Atributo	Diversidade de Valores	Combinações Possíveis/Espaço de Soluções
Carril	2	2
Carril II	2	16
Travessa	3	128
V. Máx. de Pl.	9	65 536
Comp. Total	2	262 144
Peso Total	5	8 388 608
Velocidade	4	134 217 728

Tabela 23: Espaço de soluções vs atributos antecedentes

Seleção dos Dados Através da Análise Bivariada

Com a análise bivariada também se pretende identificar correlações entre os atributos em presença, por forma a não só compreender os dados, bem como possibilitar uma simplificação do conjunto de dados tendente a centrar o algoritmo no espaço de dados mais informativo. Atendendo a que as variáveis não apresentam distribuição normal não foi aplicado o coeficiente de correlação linear de Pearson, mas o coeficiente de correlação ordinal de Spearmam, conforme anexo XVI.

Cumpre referir que a utilização deste coeficiente não permitiu efetuar o cálculo do coeficiente de correlação para os atributos qualitativos em presença, "Carril II" e "Travessa".

O anexo XVI apresenta as correlações entre todos os atributos presentes no ficheiro, assinalando-os como correlação fraca, moderada, forte e muito forte.

Validação dos Especialistas de Domínio

Conjuntamente com os especialistas de domínio foi efetuada uma validação das correlações verificadas, que se apresentam compartimentadas em três grupos:

- Correlações entre Antecedentes;
- Correlações entre Consequentes;

- Correlações ente Antecedentes e Consequentes.

Correlações entre Antecedentes

As correlações mais significativas identificadas são:

	Comprimento_Total	Peso.Total	Velocidade(Média)
Carril			0,934
Vel..Máxima.de.Planta	-0,8107	0,866	
Comprimento_Total..mediana.		-0,893	

Tabela 24: Correlações significativas entre atributos antecedentes

Neste caso verifica-se a possibilidade de existirem redundâncias entre o grupo de antecedentes.

Correlações entre Consequentes

As correlações mais significativas identificadas são:

	NIVLED1	NIVLDD1	NIVLDD2	ALIND	ALINESQR	ALINDIRR	NIVTRANS	EMPENO9M	BITMED
NIVLE	0,887								
NIVLD		0,887							
NIVLED2			0,811						
ALINE				0,951			0,853		
ALIND							0,854		
ALINED1					0,880				
ALINDD1						0,879			
ALINESQR						0,574			
EMPENO3M								0,766	
BITOLA									0,891

Tabela 25: Correlações significativas entre atributos consequentes

Neste caso verifica-se que existem atributos consequentes que pela correlação evidenciada partilham o mesmo comportamento face aos dados em presença. As correlações apresentadas podem dividir-se entre:

a. Atributos que se referem à mesma medida mas que a distância de observação do atributo varia, no caso todos os atributos referentes com o Nivelamento Longitudinal NIVLE, NIVLED1, NIVLD, NIVLDD1, e Empeno 3M, Empeno 9M ou que também referentes à mesma medida estão correlacionados pois são calculados sobre a mesma base no caso BITOLA e BITMED.

b. Atributos que se referem à mesma medida mas com componentes da estrutura de via distintas, como ALINE e ALIND (carril esquerdo e direito).

c. Atributos que se referem a medidas distintas como ALINE, ALIND e NIVTRANS, ALINED1, ALINEDD1 e ALINESQR, ALINDIRR.

O tratamento a dar a cada uma das correlações poderá ser distinto de acordo com a previsibilidade da relação. Os dois últimos grupos, B e C, são os mais informativos pois empiricamente apresentam relações menos prováveis porque versam sobre componentes diferentes e ou medidas diferentes.

Desta forma a eliminação de atributos poderá ser efetuada com maior segurança para os atributos do grupo A, sendo que para os atributos dos grupo B e C uma eventual eliminação de atributos só deve acontecer após se provar a sua redundância em sede de apreciação dos resultados de aplicação do algoritmo. Cumpre referir que com base em Andrade [7] é usual que em sede de processo de medição se despreze os atributos referentes ao alinhamento longitudinal de um dos carris (esquerdo ou direito).

Correlações entre Antecedentes e Consequentes

	Curva
ALINE	0,9033
ALIND	0,9029
NIVTRANS	0,8845

Tabela 26: Correlações significativas entre atributos antecedentes e consequentes

As três correlações apresentadas na tabela 26 são um padrão na relação entre o contexto da estrutura (curva) e o comportamento da mesma (valores dos parâmetros). Num cenário de aplicação de um algoritmo que pretenda identificar regras com uma distribuição significativamente diferente[14] da distribuição à priori, estes relações poderão não ser inferidas, logo a leitura das correlações entre antecedentes e consequentes é um auxiliar à aferição de padrões à priori e portanto complementar ao conhecimento retirado da aplicação de um algoritmo com as caraterísticas mencionadas.

Apresentando os atributos supra identificados correlações consideradas muito fortes[15], podia ser equacionada a sua substituição, contudo tal só ocorrerá se o algoritmo a aplicar apresentar morosidade. Esta decisão tem como objetivo explorar o viés indutivo do algoritmo com o maior conjunto de dados em presença, confirmando as redundâncias identificadas, ou não, em sede de análise exploratória. Contudo as correlações aqui identificadas serão objeto de atenção sempre que o algoritmo identificar um padrão.

6.5. Construção dos Dados – Ficheiro Integrado

Transformação de Dados/Discretização

A discretização é um processo que ao efetuar a transformação dos valores do atributo altera com potencial perda de informação a base de dados objeto de discretização.

O algoritmo definido por Jorge, Azevedo and Pereira [2], que permite a existência de atributos numéricos como consequente na identificação de regras de associação, reduz, face a outros algoritmos de regras de associação que obrigam à discretização de todos os atributos, a perda de informação decorrente deste processo.

[14] Este conceito, de significativamente diferente da distribuição à priori será desenvolvido no ponto Aplicação do Método.

[15] Embora não diretamente aplicável de forma geral considera-se nas ciências sociais e humanas, as correlações como muito fortes quando $|r| \geq 0.75$.

Analisado o conjunto dos atributos antecedentes verifica-se que os mais informativos são os que serão objeto de discretização, GRAD, BITOLA, INCLINAÇÃO, o que significa que o sucesso da análise na obtenção de padrões de interesse também depende do processo de discretização, desta forma colocam-se como alternativas três possibilidades para discretizar:

Seguir o critério dos autores das regras de distribuição para o exemplo real dado em Jorge, Azevedo and Pereira [2] e utilizar um método não paramétrico, supervisionado, apresentado por Fayyad and Irani [33]. Sendo um método supervisionado, este atende à informação presente nos atributos, neste caso ao critério de minimização da entropia nos intervalos resultantes da discretização. Este facto irá concorrer para minimizar um dos problemas associados à discretização em problemas de regras de associação, que é o facto de no próprio intervalo existirem comportamentos muito díspares relativamente às variáveis explicativas, conforme melhor definido em Aumann and Lindell [34]. A desvantagem desta regra é que requer que a discretização atenda ao atributo consequente, o que implicaria ter que eleger, dentro dos 20 parâmetros de via, um como atributo consequente, o que, estando à procura de relações de significância sem constrangimentos, além do suporte relativo ao antecedente e ao interesse da regra definido pela aplicação do teste KS, não se enquadra na lógica da abordagem.

Contudo, o método de discretização supra permanece como hipótese, pois poder-se-ia recorrer a especialistas para identificar entre os 20 atributos denominadas parâmetros de via quais seriam mais determinantes, e selecionando-os efetuar-se outras tantas discretizações, dando origem a igual número de bases de dados uma por cada novo elemento escolhido, obrigando o algoritmo a correr também outras tantas vezes. Com o resultado apurado, escolheríamos, através da definição de um patamar de interesse com base no teste KS, as regras com maior significância independentemente dos consequentes escolhidos.

A segunda forma de discretização possível é aplicar um modelo baseado na inspeção visual de cada um dos atributos, atendendo à análise efetuada pelos especialistas de domínio ao comportamento da distribuição dos atributos a discretizar.

A terceira forma de discretização considerada foi através do modelo de clustering K-Means, recorrendo ao R, que utiliza o algoritmo proposto por Wong [35].

O algoritmo de K-means desenvolve-se de uma forma genérica da seguinte forma:

1. Seleciona (aleatoriamente) os centroides.
2. Aloca cada um dos elementos do atributo ao centroide mais próximo.
3. Recalcula o centroide como a média de todos os centroides no cluster.
4. Aloca cada um dos elementos ao centroide mais próximo.
5. Contínua em ciclo com os passos 3 e 4 até que nenhuma observação seja reafeta a um novo cluster ou o nº máximo de iterações seja alcançado.

O algoritmo proposto por Wong [35] visa, dado um conjunto de observações $(X_1, X_2, \dots \dots . X_n)$, efetuar a partição das n observações em K conjuntos $S = \{S_1, S_2 \dots S_K\}$ de forma a minimizar a soma dos quadrados intra cluster, conforme figura infra.

$$Minimizar \sum_{i=1}^{K} \sum_{X_J \in S_I} \left\| x_j - \mu_i \right\|^2$$

Expressão (6.1): Minimização quadrática dos desvios intra cluster

O que significa que nos passos 2 e 4 as observações são alocadas ao cluster com menor valor de:

$$SQ(K) \sum_{i=1}^{n} (x_i - \mu_k)^2$$

Expressão (6.2): Cálculo do desvio intra cluster

Onde k é o cluster, i é a observação e μ o centroide do cluster.

Este é um facto a ter em conta, inserindo-se na supra mencionada perda de informação com o processo de discretização e subsequente interpretação das regras inferidas pelo algoritmo a aplicar, ou seja, atendendo ao algoritmo supra o processo de discretização não garante em termos absolutos que todas as observações encontram-se no cluster que seria suposto dado o critério da proximidade.

Apesar da advertência, face aos outros algoritmos identificados foi esta terceira opção a escolhida para efetuar a discretização, pois apresentava um bom compromisso entre a opção por um método que nos permite regular, controlar a qualidade da discretização e simultaneamente ser facilmente compreensível pelos especialistas de domínio.

Assim conforme Figura 11 efetuaram-se vários testes com o intuito de encontrar o equilíbrio entre o menor número de conjuntos e simultaneamente a minimização da soma de quadrados constante da diferença/distância (Euclidiana) de cada um dos pontos de cada conjunto relativamente ao seu centroide (média) .

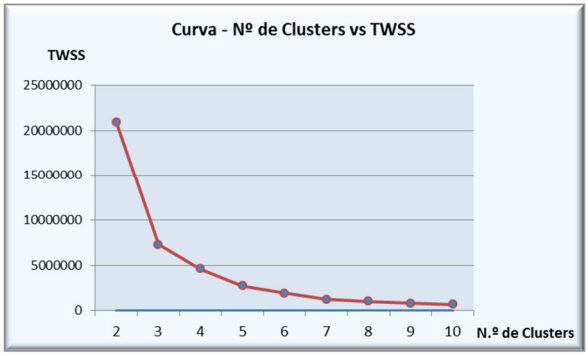


Figura 11: Nº de clusters vs TWSS – Atributo Curva

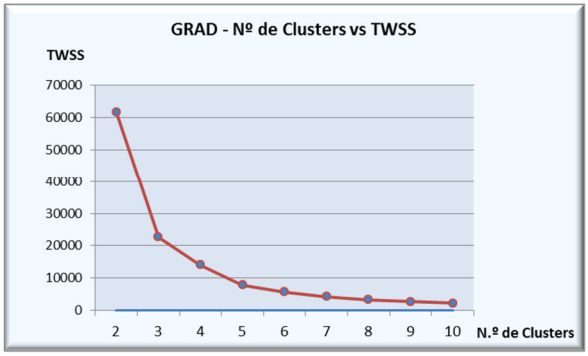


Figura 12: Nº de clusters vs TWSS – Atributo Grad

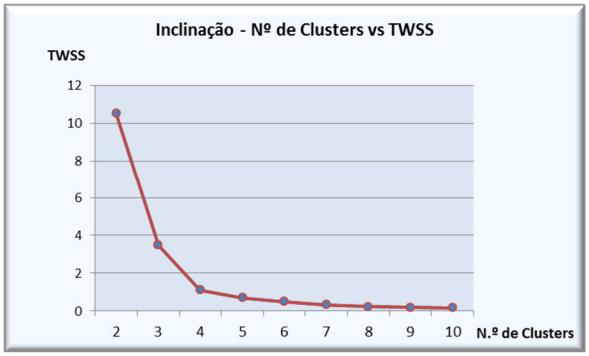


Figura 13: Nº de clusters vs TWSS – Atributo Inclinação

Como se pode verificar pelas Figuras 11,12 e 13 considerou-se que o processo de discretização teria um adequado equilíbrio entre o nº de clusters e a TWSS (soma dos quadrados intra cluster) para o valor de cinco (clusters) em todas as variáveis, pois é neste ponto que a curva que relaciona estas duas variáveis sofre uma maior redução na sua inclinação.

Cumpre referir que no processo de cálculo efetuado em R os valores da TWSS podem ser alterados se apenas fixarmos o nº de clusters, pois o processo inicia-se pela escolha aleatória dos centroides, dando origem a uma solução diferente cada vez que se inicia o processo. A forma escolhida para ultrapassar esta dificuldade foi utilizar a opção nstart, conforme expressão infra, da função Kmeans que calcula múltiplas configurações iniciais dos centroides devolvendo a melhor. Também se podia ter optado por fixar a

raiz através da função set.seed() contudo estaríamos a garantir que os resultados eram repetíveis mas não uma solução que minimizasse a TWSS.

```
kmeans(x,10,nstart=25)
```

Expressão (6.3): Expressão Kmean para cálculo dos clusters com base em múltiplos centroides

Desta forma os resultados incorporados no ficheiro final resultantes do processo de discretização para as seguintes variáveis foram:

	Cluster	Valor mais baixo	Valor mais alto	Nº Observações	% Observações
Curva	1	4,45	14,69	18 997	9%
	2	14,73	28,87	25 361	13%
	3	28,91	52,27	7 918	4%
	4	-47,66	-11,95	33 044	16%
	5	-11,91	4,41	115 377	57%
Total				**200 697**	**100%**

Tabela 27: Mapa discretização atributo “Curva"

	Cluster	Valor mais baixo	Valor mais alto	Nº Observações	% Observações
Gradiente	1	0,59	1,25	64 585	32%
	2	-2,66	-1,33	7 488	4%
	3	-0,16	0,55	70 696	35%
	4	-1,29	-0,2	22 747	11%
	5	1,29	2,54	35 181	18%
Total				**200 697**	**100%**

Tabela 28: Mapa discretização atributo “Gradiente”

	Cluster	Valor mais baixo	Valor mais alto	Nº Observações	% Observações
Inclinação	1	0,224292	0,224292	336	0%
	2	0,008809	0,019187	52 420	29%
	3	-0,005974	0,001572	71 581	40%
	4	-0,01728	-0,006833	14 313	8%
	5	0,001696	0,007496	40 265	23%
Total				**178 915**	**100%**

Tabela 29: Mapa discretização atributo “Inclinação”

7. Modelação

7.1. Seleção da Técnica de Modelação

Obtida a base de dados, na sequência do processo de conhecimento, análise e preparação dos dados, com uma dimensão de 178.915 observações, 20 atributos consequentes e 11 antecedentes conforme anexo VI, o desafio de extração de conhecimento, assenta na criação de um modelo legível e interpretável, com um algoritmo computacionalmente eficiente, que não só reforce o conhecimento empírico comummente aceite, mas que surpreenda na identificação de novas relações ou padrões de conhecimento até então desconhecidos.

Face à base de dados em apreço ser constituída na quase totalidade por atributos numéricos, só dois dos atributos são categóricos, "Carril II" e "Travessa", a escolha de um algoritmo que trate dados quantitativos de forma eficiente permitindo retirar o máximo de informação da base de dados é um objetivo.

As técnicas de *Data Mining* aplicadas com maior frequência ao tratamento de dados numéricos/quantitativos dividem-se entre as técnicas de regressão (árvores de regressão, redes neuronais,..) e as regras de associação com prévia discretização Webb [36].

A vantagem da aplicação das regras de associação face às técnicas de regressão é a da sua completude, pois elas permitem encontrar todas as relações existentes na base de dados apenas dependendo do mínimo suporte e confiança definido pelo utilizador do método.

A desvantagem das regras de associação face às regras de distribuição, é que as primeiras apenas estabelecem um valor para a variável consequente e obrigam a uma prévia discretização da mesma, correndo o risco de perder-se parte da informação a ela associada.

Desta forma é selecionado o método para a modelação de dados proposto por Jorge, Azevedo and Pereira [2], denominado regras de distribuição, neste, conforme supra exposto, é mantida a vantagem das regras de associação, ou seja, permite encontrar todas as relações existentes evitando em simultâneo o processo de discretização e subsequente perda de informação, na medida em que apresenta toda a distribuição do consequente dadas as variáveis antecedentes. É com base nesta diferença, ou seja na existência de uma distribuição para o consequente, que o processo de seleção de regras é conduzido, pois, após definido o suporte mínimo para os antecedentes, o processo de seleção de regras baseia-se na diferença da distribuição de cada consequente dado os antecedentes em presença e a sua distribuição à priori.

Conforme definido em Jorge, Azevedo and Pereira [2] uma regra de distribuição é uma regra A=>y=$D_{y/A}$, onde A é um conjunto items, y é a variável de interesse e $D_{y/A}$ é a distribuição empírica de y para todos os casos em que A ocorre.

A tarefa de descoberta de regras de distribuição consiste em encontrar todas as regras de distribuição do tipo A=>y=$D_{y/A}$, em que A cumpre com o suporte mínimo definido e $D_{y/A}$ é significativamente diferente do ponto de vista estatístico (face a um limiar pré-definido) da distribuição associada a $D_{y/A}$, para o universo de valores da base de dados.

Acresce referir que de acordo Srikant and Agrawal [37] também existem desvantagens destes métodos, regras de associação e regras de distribuição, face aos métodos de regressão, estas derivam sobretudo da quantidade de informação que pode ser selecionada pelo modelo para observação pelo utilizador

Após a definição da base de dados e da escolha do método, inicia-se o processo da sua aplicação.

7.2. Aplicação do Algoritmo

7.2.1. Revisão da Literatura/Enquadramento Teórico do Algoritmo

Os algoritmos de regras de associação, no contexto de dados quantitativos, são abordados pela primeira vez por Piateski and Frawley [38], em que é apresentado um algoritmo bastante simples que consistia em marcar um dos valores do atributo e efetuar uma contagem para toda a base de dados dos valores dos outros atributos que coincidiam na mesma transação com o valor marcado, o resumo era utilizado para derivar as regras de associação. O algoritmo tinha que correr uma vez por cada atributo e, para se encontrar todas as regras subjacentes à base de dados, tinha que se guardar um resumo de todos os resumos produzidos pelo algoritmo para todas as combinações de atributos, o que tornava a procura exponencialmente extensa.

Srikant and Agrawal [37] forneceram a primeira definição geral e algoritmo para tratamento deste caso específico de regras de associação. Para que se tenha uma noção da especificidade das regras de associação para dados quantitativos, recorda-se que o primeiro algoritmo para a extração de items sets frequentes e regras de associação, denominando Apriori, foi publicado em 1994 por and and Srikant [39]. Só dois anos mais tarde, o mesmo autor em Srikant and Agrawal [37] publica a sua abordagem a dados contínuos. Este hiato prende-se com o contexto inicial das regras de associação, que estavam direcionadas para trabalhar com *item-sets* discretos no contexto de grandes superfícies de retalho. Neste trabalho de Srikant and Agrawal [37], a abordagem consiste em efetuar um mapeamento dos atributos contínuos para intervalos ou valores categóricos através de um processo de discretização. Só depois, usando o algoritmo sobre estes valores discretizados, o algoritmo retorna todas as regras de associação, sendo-lhe aplicado um filtro para eliminar as menos significativas.

Fukuda, Morimoto, Morishita and Tokuyama [40] apresentam um algoritmo eficiente para encontrar regras de associação dado o suporte e a confiança. Contudo, direciona-se para os casos em que o consequente é categórico. Fukuda, Morimoto, Morishita and

Tokuyama [40] também dão um contributo para a otimização do suporte das regras, com vista a ter um retorno de regras mais selecionado e mais rápido.

Morimoto, Ishii and Morishita [41], no âmbito das árvores de regressão, propõem algoritmos para espartilhar o espaço de soluções, composto por atributos contínuos, em regiões ótimas.

Zhang, Lu and Zhang [42] dão um contributo para o processo de discretização de atributos quantitativos, apresentado em Srikant and Agrawal [37], através da utilização de *clustering* para otimizar as partições.

Fukuda, Morimoto, Morishita and Tokuyama [43], Fukuda, Morimoto, Morishita and Tokuyama [40] e Yoda [44] apresentam o caso em que dois atributos numéricos ocorrem no consequente.

Relativamente a uma abordagem estatística no âmbito dos processos de regras de associação, temos um trabalho de Brin, Motwani, Ullman and Tsur [45] que utiliza pela primeira vez os métodos de inferência estatística para confirmar a adequabilidade das regras resultantes do algoritmo.

Aumann and Lindell [34], no âmbito da apresentação de um processo de discretização que permite quantificar a informação perdida, propõem um teste estatístico Z que mede o interesse da regra face aos resultados apresentados pelo algoritmo.

Webb [36], no desenvolvimento do trabalho efetuado por Aumann and Lindell [34], discorre sobre a adequabilidade do teste estatístico Z proposto, e sugere um teste T, também para efeitos de validação do interesse das regras geradas pelos algoritmos.

Assim, chegamos ao trabalho de Jorge, Azevedo and Pereira [2], que propõe o algoritmo subjacente a esta dissertação.

É oportuno referir que o conceito de regras de distribuição, e o seu potencial associado, reside no tratamento dos dados em bruto dos consequentes. Este facto diferencia-as, conforme mencionado pelos autores Jorge, Azevedo and Pereira [2], das técnicas de regras de associação que implicavam a discretização, e da técnica proposta por Aumann and Lindell [34] que propunha a substituição dos valores dos consequentes pelos valores da sua média ou mediana.

Conforme definido em Jorge, Azevedo and Pereira [2], as regras de distribuição podem ser utilizadas para tarefas de descoberta de padrões, bem como para tarefas preditivas. Definem, como *tresholds* para a escolha das regras, o suporte e o interesse baseado no teste Kolmogorov Smirnov, este dá uma medida de dissemelhança entre a distribuição estatística do consequente, face aos antecedentes em presença e a distribuição á priori daquele consequente, que funciona como um filtro para a melhoria dos resultados obtidos, de forma a selecionar ainda mais as regras que se pretendem devolvidas pelos modelo.

7.2.2. Aplicação do Algoritmo/Construção do Modelo

Preparação Para a Aplicação do Algoritmo

Para a aplicação do algoritmo de regras de distribuição proposto por Jorge, Azevedo and Pereira [2] terá que ser dado como entrada uma base de dados DB, um valor para o suporte mínimo das regras (minsup) e um valor para o interesse das regras α.

Base de Dados

A base de dados fornecida ao algoritmo para obtenção das regras de distribuição foi a constante do ficheiro integrado representado no anexo VI (“Depois da Eliminação Manual de Atributos”).

Definição dos Parâmetros do Modelo

Os parâmetros de calibração do modelo definem-se pelo suporte mínimo das regras (minsup), que representa o número de observações/túpulas associada a cada regra

gerada, e pelo valor α associado ao teste de Kolmogorov Smirnov (KS), critério de interesse da regra.

A calibração destes parâmetros pode ser definida atendendo ao aspeto quantitativo, número de regras geradas, e ao aspeto qualitativo, definindo-se um critério de qualidade das regras através de um parâmetro objetivo ou recorrendo aos especialistas de domínio.

Calibração dos Parâmetros – Aspeto Quantitativo

Conforme Tabela 30 e 31 estes dois critérios, minsup e valor α, foram testados para a base de dados em apreço, circunscrito a uma amostra do conjunto de variáveis, sendo que, para os valores testados, apenas a alteração do suporte da regra (minsup) revelou condicionar de forma significativa o número de regras geradas.

					Nº de Regras por Parâmetro				
Teste	α	Suporte	Valor Absoluto do Suporte	Items Frequentes	BITOLA	NIVLD	NIVLE	NIVLED1	ALINED1
1	0,01	0,005	895	87	2 396	1 568	1 215	1 311	1 443
2		0,05	8 946	36	154	121	116	121	120
3		0,1	17 892	28	65	42	40	45	47
4		0,15	26 838	26	47	33	29	31	36
5		0,2	35 783	20	15	10	8	8	12
6		0,3	53 675	13	8	4	4	4	7

Tabela 30: Suporte da regra vs nº de regras geradas

					Nº de Regras por Parâmetro				
Teste	α	Suporte	Valor Absoluto do Suporte	Items Frequentes	BITOLA	NIVLD	NIVLE	NIVLED1	ALINED1
1	0.05	0,1	17 892	28	65	46	45	49	50
2	0,01		17 892	28	65	42	40	45	47

Tabela 31: Valor de significância do teste KS vs nº de regras geradas

De acordo com os resultados produzidos foi selecionado para a aplicação do algoritmo a todos os atributos, o valor de 0.1 para o parâmetro minsup, e um α de 0.01. Estes valores garantem a produção pelo algoritmo de um número de regras, que se entendeu como o limite para uma base de trabalho destinada à inspeção visual.

Calibração dos Parâmetros – Aspeto Qualitativo

Para a calibração dos parâmetros, no que respeita ao aspeto qualitativo das regras produzidas, recorreu-se a um critério objetivo suportado no indicador estatístico que é utlizado na avaliação dos parâmetros de via, ou seja como o "negócio" define o seu critério de qualidade. Desta forma foi selecionado o desvio padrão associado às distribuições subjacentes a cada uma das regras produzidas e efetuada uma análise e seleção dos valores dos parâmetros com base na amplitude do desvio padrão produzido pelas regras geradas, assegurada a não existência de outliers.

			BITOLA	**NIVLD**	**NIVLE**	**NIVLED1**	
Teste	α	Suporte	Amplitude	Amplitude	Amplitude	Amplitude	Amplitude
1	0,01	0,005	10,44	4	3,71	2,39	3,36
2		0,05	6,4	2,47	2,21	1,43	1,61
3		0,1	6,22	2,19	1,93	1,38	1,61
5		0,2	5	1,06	0,98	0,68	1,08
6		0,3	4,34	0,49	0,44	0,67	1

Tabela 32: Alteração do suporte da regra vs amplitude dos desvios padrões das regras afetas a cada variável

			BITOLA	**NIVLD**	**NIVLE**	**NIVLED1**	
Teste	α	Suporte	Amplitude	Amplitude	Amplitude	Amplitude	Amplitude
1	0.05	0,1	6,22	2,38	2,098	1,38	1,61
2	0,01		6,22	2,19	1,93	1,38	1,61

Tabela 33: Valor de significância do Teste KS vs amplitude dos desvios padrões das regras afetas a cada variável

Conforme Tabela 32 e 33 conclui-se que o parâmetro minsup, que define o suporte da regra, é o que mais contribui para a variação da amplitude dos desvios das regras geradas.

Desta forma tendo em análise os resultados produzidos no quadro supra, é para o valor de 0.005, referente ao parâmetro minsup, que obtemos uma maior amplitude, pelo que se atendêssemos apenas ao aspeto qualitativo das regras seria este o valor escolhido.

Contudo quando atendemos ao nº de regras geradas pelo valor selecionado em sede de calibração quantitativa, 0,005, este situa-se sempre acima das 1000 regras geradas tornando impraticável a leitura das mesmas, pelo que se opta pelo valor de 0,1 para o parâmetro minsup, e um α de 0,01 para a aplicação do algoritmo a todos os atributos.

Aplicação do Algoritmo

O algoritmo foi aplicado para o conjunto das 20 variáveis de interesse, consequentes, 11 antecedentes e com o valor de parâmetros supra definido.

No anexo XVII encontramos um conjunto de exemplos das regras geradas e respetivas distribuições mais significativas, ou seja com maior suporte e maior afastamento ao desvio padrão da distribuição à priori. Como podemos visualizar a distribuição à priori encontra-se representada nos gráfico a cinzento claro e a distribuição associada à regra gerada encontra-se a cor preta, o(s) antecedente(s) encontram-se a cor vermelha fora do gráfico, a azul os valores dos parâmetros, a média e o desvio padrão, a preto o consequente.

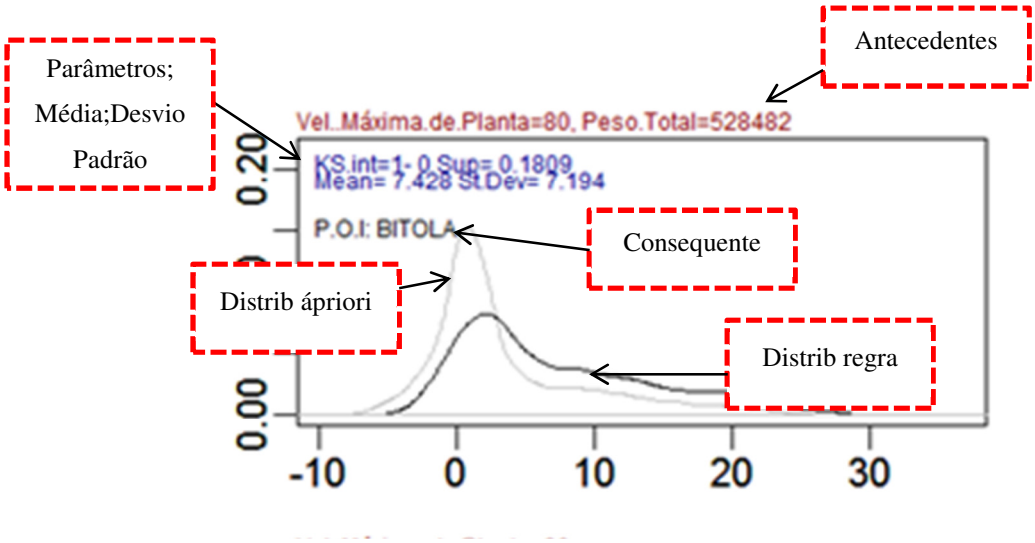


Figura 14: Representação regra de distribuição

Conforme também se constata no anexo XVII, existem quadros com circulo a cor vermelha e outros a cor verde, estes assinalam respetivamente a descoberta de padrões com pior e melhor comportamento à luz dos parâmetros definidos para o negócio.

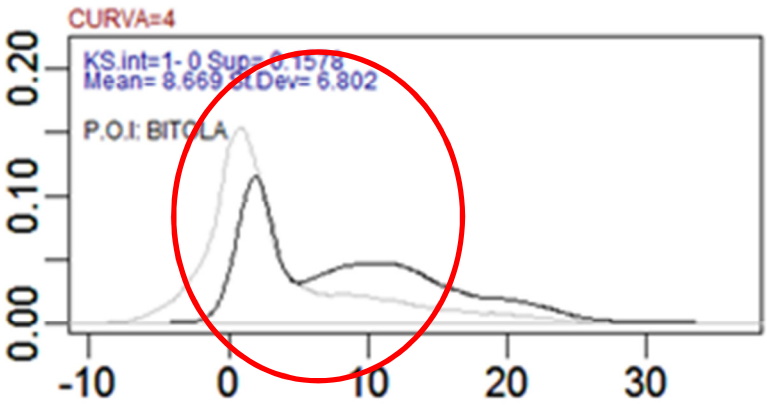


Figura 15: Regra com distribuição incorreta de acordo com os parâmetros do negócio

O padrão supra exposto da regra Curva 4 => Bitola (com uma distribuição estatisticamente diferente, segundo o teste de KS para um valor de $\alpha = 0{,}01$, da distribuição à priori do atributo bitola) mostra uma distribuição mais achatada e com maior cauda que a distribuição à priori (a cinza claro). Para além deste facto, contrariamente à distribuição à priori, a distribuição da regra gerada não tem a maior frequência das suas observações no valor zero, que é o valor ideal para avaliação do parâmetro bitola. Desta forma estamos em presença de uma distribuição com pior comportamento à luz dos parâmetros de qualidade do negócio, pois têm um desvio padrão superior e em que a maior frequência das suas observações não está no valor de referência dado pelos especialistas de negócio e confirmado pela distribuição à priori.

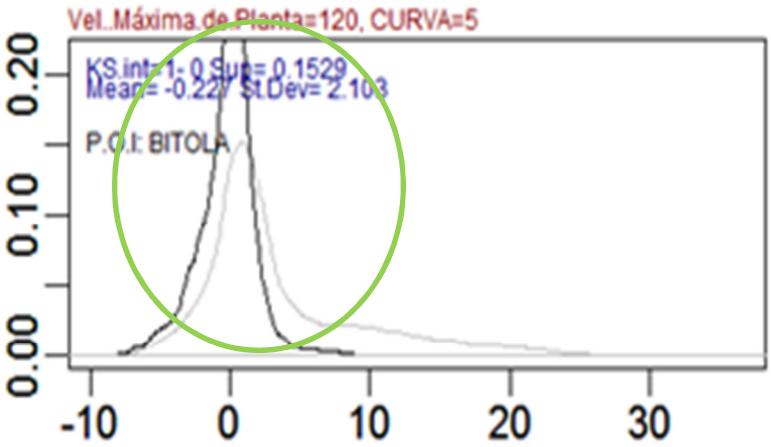


Figura 16: Regra com distribuição correta de acordo com os parâmetros do negócio

De forma oposta à avaliação efetuada ao quadro anterior, neste figura o padrão resultante da regra Vel. Máxima de Planta = 120 & Curva 5 => Bitola (com uma distribuição estatisticamente diferente segundo o teste de KS para um valor de α = 0,01 da distribuição à priori do atributo bitola) mostra uma distribuição mais pontiaguda e concentrada que a distribuição à priori (a cinza claro), o que significa uma distribuição com melhor comportamento, à luz dos parâmetros definidos pelo negócio, do que a distribuição à priori.

Estes dois casos mostram que a identificação de padrões anómalos não significa que sejam padrões com uma conotação negativa no enquadramento dado pelo contexto de negócio.

7.3. Avaliar o Modelo

Conforme metodologia CRISP-DM Chapman [4] esta avaliação é referente à modelação do modelo circunscrita a fatores como a capacidade de generalização e precisão.

Não se tratando de um algoritmo tendente à classificação, a precisão do modelo não tem uma medição objetiva para além da aferição relativa à correção na identificação de

distribuições estatisticamente diferentes das distribuições à priori. Esta é garantida pelo teste estatístico de KS com um α de 0,01.

Relativamente à generalização do modelo, a análise efetuada pelo especialista de domínio aos resultados da aplicação do algoritmo referidos no ponto anterior, conduziu à perceção de que uma parte significativa das regras aí geradas poderiam relacionar-se com subconjuntos de dados dentro do conjunto total, estes subconjuntos correspondiam a cada um dos cinco troços em presença.

A importância desta descoberta é que cerceava a capacidade de generalização das regras que apresentassem as caraterísticas supra expostas, podendo contudo dar uma oportunidade de classificar os troços através dos padrões gerados, uma espécie de metadados relativamente à classificação das componentes de via expostas nos atributos antecedentes.

Desta forma efetuou-se uma análise às caraterísticas de cada um dos troços, tendo-se identificado características únicas de cada troço que sempre que presentes numa regra identifica-a como pertencente a determinado troço e apenas a esse.

Atributos Caraterizadores dos Troços*							
Troço	**Carril**	**Carril II**	**Travessa**	**Vel Max de Planta**	**Comprimento Total (Mediana)**	**Peso Total**	**Velocidade (Média)**
Troço E	54	Barra Curta	Madeira	75;80;90;100;110	70	528 482	113
Troço A	60	BLS	Betão Monobloco	80;100;110;120	66,8	1 541 135	150
Troço C	54	BLS	Bibloco	95;110;120;125;140	66,8	6 271 537	150
Troço B	54	BLS	Bibloco	140	66,8	7 683 468	151
Troço D	54	Barra Curta	Madeira	120	66,8	7 778 428	146

* Acresce o atributo Gradiente;Curva e Inclinação.

Tabela 34: Atributos caraterizadores dos troços

Após a identificação referida, classificou-se cada uma das regras geradas para cada uma dos atributos identificando as regras que apenas pertenciam a cada um dos troços. Por forma a reduzir o tempo afeto a esta reclassificação e concluindo que parte das correlações identificadas em sede da secção 6.5 verificaram-se em sede de produção de

regras excluíram-se os atributos que apresentavam correlações nas regras geradas, conforme é exemplo a figura infra que identifica visualmente as correlações entre as distribuições dos atributos NIVLE (os 9 gráficos do lado esquerdo) e NIVLD.

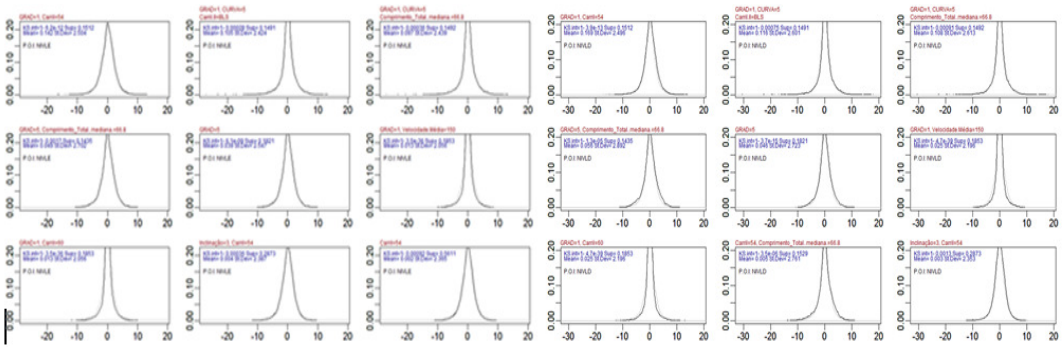


Figura 17: Correlação entre os atributos "NIVLE" e "NIVLD" em sede de produção de regras

Tratando-se de classificação de troços excluíram-se também os atributos que não se encontravam presentes em todos os troços[16], ficando a aplicação do algoritmo definida para os seguintes atributos:

	NIVLE	NIVLD	NIVLED1	NIVLDD1	NIVLED2	NIVLDD2	ALINE	ALIND	ALINED1	ALINDD1
Correlação	X	X						X		
Não Aplicável					X	X				
Atributo Selecionado			X	X			X		X	X

	ALINESQR	ALINDIRR	NIVTRANS	NIVTRANR	EMPENO3M	EMPREL3M	EMPENO9M	EMPREL9M	BITO1A	BITMED
Correlação							X		X	
Não Aplicável										
Atributo Selecionado	X	X	X	X	X	X		X		X

Tabela 35: Seleção de atributos pós regras

Desta forma das 637 regras geradas 40% foram classificadas como de um só troço, distribuídas da seguinte forma:

[16] Os atributos terminados em D2 referem-se a monitorização de parâmetros para velocidades superiores a 160 km/h estas velocidades só são praticadas em alguns troços em presença.

Troço	α	Suporte	Nº de Regras	% de Regras	Nº de Observações	% de Observações
Troço E	0,01	0,1	139	0	73 027	41%
Troço A			100	0	43 492	24%
Troço C			17	0	35 040	20%
Troço B			-	-	14 149	8%
Troço D			-	-	13 207	7%
Não Afetas a Nenhum Troço			381	1	0	0%
TOTAL			637	1	178 915	100%

Tabela 36: Distribuição de regras por troço

Compete referir que as regras remanescentes, não diretamente afetas por nenhum atributo a nenhum troço, encontravam uma possibilidade de, quando explorada a sua composição, pertencerem na sua maior parte a um determinado troço. Exemplo deste facto é a regra infra.

Ant_sup	Stdev	Dist	Subgroup
0,188363	3,36073	EMPENO3M	c("GRAD=3", "Inclinação=3", "Travessa=Madeira")

Figura 18: Regra com "falsa" capacidade de generalização

Esta regra apesar de não ter nenhum atributo diretamente afeto a um troço, se atendermos à Tabela 36 que documenta as caraterísticas dos troços, visualizamos que só existem dois troços com atributo "Travessa = Madeira", são os Troços E e D.

Se atendermos às Tabelas 37 e 38, que documentam a composição dos atributos "Gradiente" e "Inclinação", verificamos que o troço D tem no máximo 2%[17] de 62.423 observações (total das observações para o cluster 3 do atributo Gradiente) o que perfaz 1097 observações.

[17] Tratando-se a regra de uma conjunção de atributos o valor menor de cada troço presente nos atributos determina o valor máximo da partição do conjunto para determinado troço.

	Cluster	Nº de Observações	Troço C	Troço B	Troço A	Troço E	Troço D
	1	336	100%	0%	0%	0%	0%
	2	52 420	24%	0%	48%	13%	16%
Inclinação	3	71 581	20%	10%	8%	57%	4%
	4	14 313	6%	0%	42%	38%	13%
	5	40 265	18%	17%	16%	49%	0%
Total		178 915	20%	8%	24%	41%	7%

Tabela 37: Composição do atributo "Inclinação"

	Cluster	Nº de Observações	Troço C	Troço B	Troço A	Troço E	Troço D
	1	60 201	20%	10%	35%	34%	1%
	2	6 224	0%	0%	97%	2%	1%
Gradiente	3	62 423	13%	6%	15%	64%	2%
	4	17 480	2%	13%	40%	30%	16%
	5	32 587	45%	7%	0%	21%	27%
Total		178 915	20%	8%	24%	41%	7%

Tabela 38: Composição do atributo "Gradiente"

Se atendermos ao suporte da regra em apreço o valor é de 0.19 o que totaliza 33.993 observações, o que se dividirmos as 1097 observações pelo valor de 33.993, permite-nos afirmar que a participação do Troço D na regra é de apenas 3% das observações que a suportam sendo a participação do Troço E de 97%.

Repartição da Regra por Troço		
Troço	**Nº de observações**	**% de observações**
D	1.097	3%
E	32.897	97%
Total da Regra	**33.994**	**100%**

Tabela 39: Repartição da regra (Gradiente =3; Inclinação=3;Travessa= Madeira => Empeno 3m com Distribuição y) por troço

Os factos apresentados inviabilizam, sem trabalho subsequente, a generalização que pretendia que os componentes de via fossem os únicos elementos explicativos da degradação dos parâmetros de via, ou seja que em face daquele antecedente,

generalizável para todo o contexto da infraestrutura de via, obteríamos determinada distribuição associada ao consequente, este facto deixa de ser verdade, pois só acontece em presença do subconjunto correspondente ao troço x.

Assim, este conjunto de dados foi explorado com o intuito de deteção de padrões para caraterização do troço, tendo sido produzido o seguinte *output*:

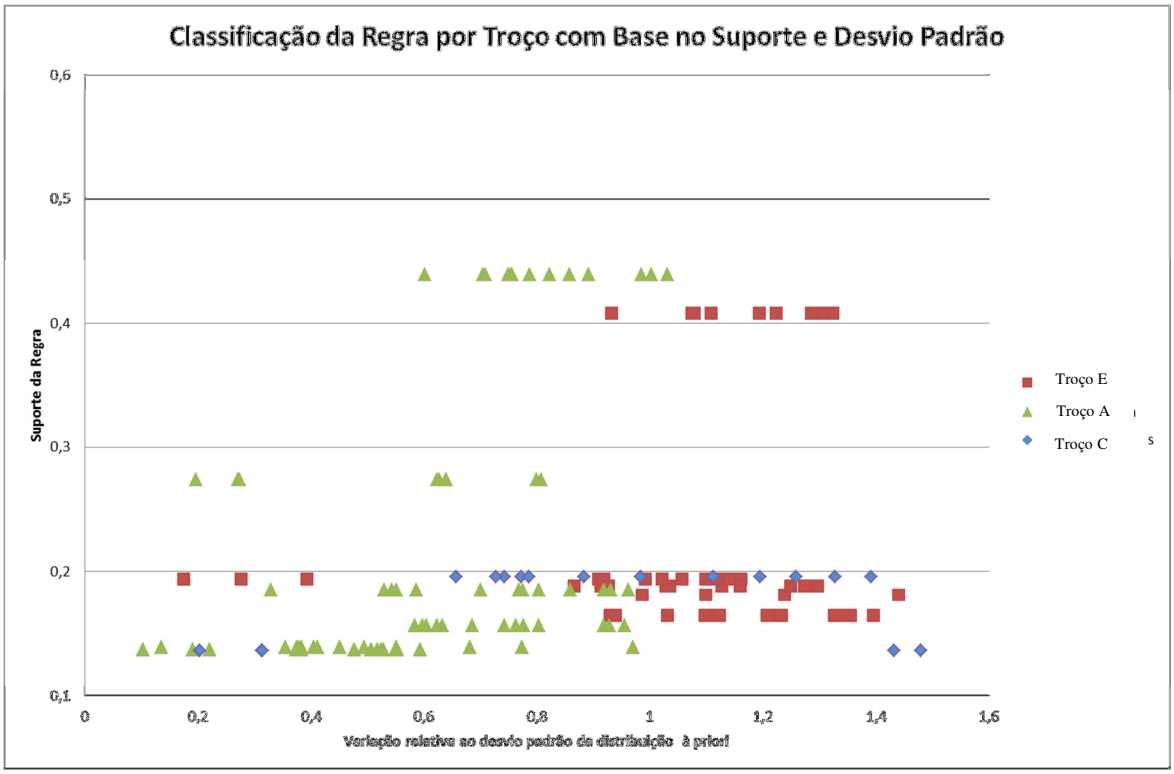


Figura 19: *Output* de visualização de deteção de padrões

Este gráfico visa obter padrões, através da classificação por troço de todas as regras relativas a todos os parâmetros. Desta forma foram definidos dos eixos de classificação para cada uma das regras, o eixo das ordenadas indicando o suporte da regra e o eixo das abcissas indicando a variação do desvio padrão face ao desvio padrão da distribuição à priori em cada uma das regras. Esta matéria será tratada no ponto seguinte avaliação.

8. Avaliação

A diferença deste ponto para o anterior reside em que neste ponto não se avalia o modelo na aceção técnica mas na dimensão do negócio, em que medida é que o modelo implementado consegue cumprir o definido no final do ponto 2.1 objetivos do negócio, ou seja em que medida é que o modelo consegue identificar uma ou mais regras/padrões interessantes.

Sobre o interesse da regra cabe referir :
Piateski and Frawley [38] definem três princípios a que as medidas de interesse (M) devem obedecer:

- Se A e B são duas variáveis estatisticamente independentes, então P(A,B) = P(A).P(B) e M=0
- A medida de interesse cresce se P(A,B) crescer, e *ceteris parabus*.
- A medida de interesse decresce se P(A), ou P(B), decrescerem, e *ceteris parabus*.

Estas medidas são expressas pelas denominadas medidas lift e convição. O teste de qui quadrado também é regularmente usado para testar a independência.

No caso das regras de distribuição, o interesse da regra pode ser medido pela diferença entre a distribuição do consequente, dados os valores apresentados nos item-sets do antecedente e a distribuição do mesmo consequente dada toda a população. Jorge, Azevedo and Pereira [2] socorrem-se do teste de Kolmogorov Smirnov (KS) como medida de diferença entre as duas distribuições a comparar.

Tal como referido por Jorge, Azevedo and Pereira [2], a aplicação para efeitos de significância das regras do teste t como do teste z, conforme adotados por Aumann and Lindell [34] e Webb [36], respetivamente, não é adequada, pois estes assumem a normalidade das distribuições que, na prática, não pode ser garantida. Desta forma, a aplicação do teste KS é a adotada para avaliar o interesse da regras.

Definição:
Dado um conjunto de transações DS, uma variável de interesse y em DS e uma regra de distribuição A=>$D_{y/A}$ obtida de DS, o valor de interesse definido por KS para essa regra é 1-p onde p é o p-value do teste de KS para as duas distribuições $D_{y/A}$ e $D_{y/\theta}$

Contudo conforme citado por Silberschatz and Tuzhilin [46], tem sido observado que as medidas objetivas de interesse da regra, como as que acima se definem, não captam toda a complexidade do processo de descoberta de padrões e que é necessário recorrer às medidas subjetivas de interesse para definirmos um padrão como interessante. Define que as medidas subjetivas de interesse não dependem apenas da estrutura da regra e dos dados utilizados na aplicação do algoritmo mas também do especialista de domínio que avalia o padrão.

Silberschatz and Tuzhilin [46] propõem uma classificação de medidas de interesse subjetivas independentes do contexto e identifica duas ordens de razão pela qual um padrão subjetivo é classificado como interessante, por o padrão ser acionável e por ser inesperado. Definem acionável através de uma noção de senso comum embora com um exemplo, contudo a definição de inesperado é objeto de uma abordagem sistemática.

Para Silberschatz and Tuzhilin [46] um padrão é inesperado à luz do sistema de crenças do avaliador, e refere que existem duas grandes abordagens à definição de um sistema de crenças:

- A publicada por C. Alchourron [47] que é uma visão absoluta ou se acredita ou não, na qual o sistema de crenças funciona da seguinte forma:

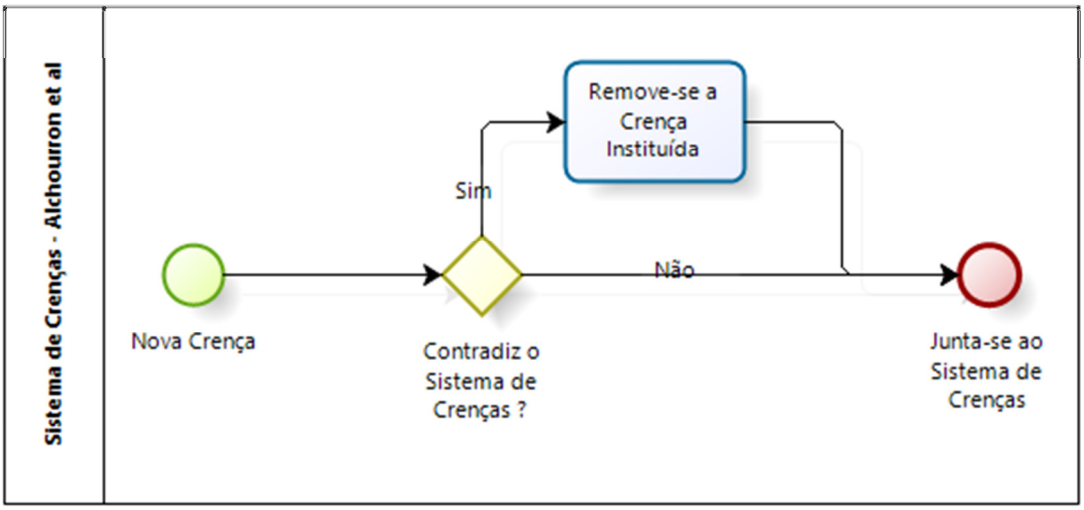


Figura 20: Sistema de crenças Alchouron et al.

- A segunda que é uma abordagem parcial por oposição à visão absoluta, traduz-se na possibilidade de se poder acreditar apenas parcialmente em determinada regra, esta é traduzida na prática por uma abordagem bayesiana em que a maior ou menor crença se traduz na probabilidade condicional associada a evidências prévias.

Silberschatz and Tuzhilin [46] adotam a segunda abordagem e pressupõem que existem dois tipos de crença, crenças *soft* e crenças *hard* :

- Crenças *soft* são aquelas que o utilizador está disposto a alterar à medida que vão sendo descobertos novos padrões baseados em novas evidências. É atribuído um grau a cada crença que expressa a confiança nessa crença. Isto significa que se α é uma crença e Ɛ a evidência prévia dessa crença:

$$Grau\ de\ crença\ em\ \alpha = P\ (\alpha/\mathcal{E})$$

Expressão (8.1): Grau de crença

- Crenças *hard* são restrições que não são alteradas com novas evidências, o especialista de domínio que avalia as regras nunca muda as crenças deste tipo.

Desta forma Silberschatz and Tuzhilin [46] classifica o interesse de um padrão/regra face ao sistema de crenças definindo que:

- Crenças *hard* - Se um padrão contradiz o sistema de crenças hard, este padrão é sempre importante e como o sistema não pode ser alterado, é pois uma indicação que os dados não estão corretos.
- Crenças *soft* – Relativamente a este tipo de crenças é proposto uma definição formal de um interesse de um padrão þ relativamente a um sistema de crenças (*soft*) β, dada por:

$$I(\text{þ}, \beta) = \sum_{\alpha \in \beta} \frac{|P(\alpha|\text{þ}, \text{Ɛ}) - \text{P}(\alpha|\text{Ɛ})|}{P(\alpha|\text{Ɛ})}$$

Expressão (8.2): Classificação do interesse da regra face ao sistema de crenças instituído

O somatório acima define o quanto um novo padrão þ altera o grau do sistema de crenças *soft*, tornando-se o padrão tanto mais interessante quanto maior for a alteração dada, ou seja quanto mais for inesperado.

O corolário desta classificação é a definição de um teorema em Silberschatz and Tuzhilin [46] que formaliza à luz da definição da expressão 8.2 que padrões inesperados são mais interessantes.

Face ao definido Silberschatz and Tuzhilin [46] sintetizam as ações a tomar face à abordagem proposta sempre que um novo padrão é descoberto.

		Padrões	
		Contraditórios	Não Contraditórios
Crenças	hard	Alterar os Dados	Aceitar os Dados
	soft	Verificar os Dados	Aceitar os Dados

Tabela 40: Descoberta de novos padrões vs sistema de crenças - ações a implementar com os dados em presença

		Padrões	
		Contraditórios	Não Contraditórios
Crenças	hard	Não alterar Sist. Crenças	Não alterar Sist. Crenças
	soft	Condicionada à Verificação dos Dados	Atualizar o Grau das Crenças do Sistema

Tabela 41: Descoberta de novos padrões vs sistema de crenças - ações a implementar com o sistema de crenças

Por fim Silberschatz and Tuzhilin [46] indicam um conjunto de questões que sintetizam o contexto da definição das medidas de interesse subjetivas, sendo as mais pertinentes:

- Como formalizar o conceito de acionável e perceber melhor como se relaciona com o conceito de inesperado;
- Como as medidas subjetivas e objetivas podem ser combinadas numa única medida integradora;
- Como calcular o interesse de um padrão relativamente a um sistema de crenças;
- Como deve ser definida a estrutura de um sistema de crenças;
- Como manter um sistema de crenças operacional;
- Como formalizar a relação entre grau de crença e medida de evidência (abordagem Bayesiana).

Mcgarry [48] no seu survey de medidas de interesse para a descoberta de conhecimento, define a seguinte taxonomia de medidas de interesse:

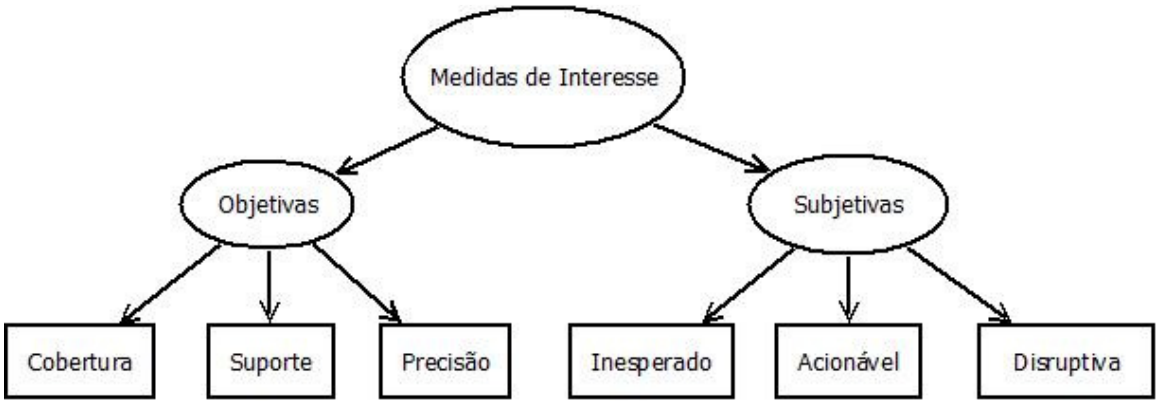


Figura 21: Taxonomia de medidas de interesse

Relativamente aos conceitos definidos por Silberschatz and Tuzhilin [46], já referenciados, é acrescentado o conceito de "disruptivo" o qual é definido por Ludwig [49] como, um padrão H é disruptivo face a um conjunto de crenças B se H não for derivável de B. Assim se um padrão contradiz o sistema de crenças este deve ser surpreendente e disruptivo.

Com o enquadramento supra definido para a classificação do interesse de padrões foram definidas as seguintes medidas tendentes à classificação dos padrões regras produzidas pelo algoritmo no ponto 7.3. .

Medidas Objetivas de Interesse

Assim no algoritmo regras de distribuição existem duas medidas objetivas do interesse da regra, o suporte e a diferença dada pelo Teste de KS para as duas distribuições em comparação (da regra e da distribuição à priori), expressa pelo p.value.

Contudo esta última medida de interesse não permite discriminar o conjunto de regras em apreço pois o valor associado ao parâmetro que define a significância do teste KS foi estabelecido em 0,01, o que devolve um conjunto de regras com uma diferença entre si muito próxima de zero.

Sendo o desvio padrão um critério de negócio sobre a avaliação da qualidade dos parâmetros de via, contendo cada uma das distribuições geradas pelo algoritmo um desvio padrão, definiu-se o mesmo como critério de avaliação do interesse da regra.

Desta forma considerou-se que o interesse da regra tinha uma relação positiva com o incremento da distância do desvio padrão da regra face ao desvio padrão da distribuição à priori.

Para que fosse possível comparar desvios padrão entre as diferentes variáveis em apreço, tendo estas diferentes ordens de grandeza no que concerne ao desvio padrão, foi necessário efetuar um processo de normalização atendendo à distribuição à priori de cada atributo. Assim criou-se uma nova medida de interesse γ que expressa através do desvio padrão a importância da regra, permitindo a comparação entre as diferentes variáveis.

Variável A		
Regra	**Std Dev**	**γ =|1-(Std Dev Regra/Std Dev À priori)|**
Regra AA	3,7906	0,47
Regra AB	2,5276	0,02
À priori	2,5745	-

Variável Z		
Regra	**Std Dev**	**γ =|1-(Std Dev Regra/Std Dev À priori)|**
Regra ZA	6,2066	0,17
Regra ZB	0,5474	0,90
À priori	5,3011	-

Hierarquização da Importância das Regras Conforme Desvio Padrão		
Regra	**γ**	**Ordem**
Regra ZB	0,9	1
Regra AA	0,47	2
Regra ZA	0,17	3
Regra AB	0,02	4

Tabela 42: Processo de normalização do desvio padrão para comparação entre variáveis

Neste ponto concluímos as medidas objetivas de interesse, definidas pelo suporte e o desvio padrão normalizado expresso pela medida de interesse γ.

Medidas Subjetivas de Interesse

Estas medidas são decompostas conforme Figura 21 em inesperadas. acionáveis e disruptivas.

Relativamente ao acionável conforme Silberschatz and Tuzhilin [46], é dependente não apenas do domínio do problema mas dos objetivos do especialista de domínio/organização num determinado momento do tempo. Desta forma não foi possível definir em tempo útil um quadro que permitisse classificar cada uma das regras pela sua acionabilidade.

Relativamente ao inesperado e disruptivo não sendo possível aplicar um modelo supervisionado, pois não existe classificação das observações em qualquer categoria, não é possível utilizar a abordagem Bayesiana proposta por [Silberschatz and Tuzhilin [46]], que permitisse qualificar as regras como interessantes. Desta forma recorreu-se a técnicas de visualização para a aferição subjetiva de padrões interessantes, técnica bastante comum ao tratamento de problemas de regras de associação.

Análise do Interesse das Regras

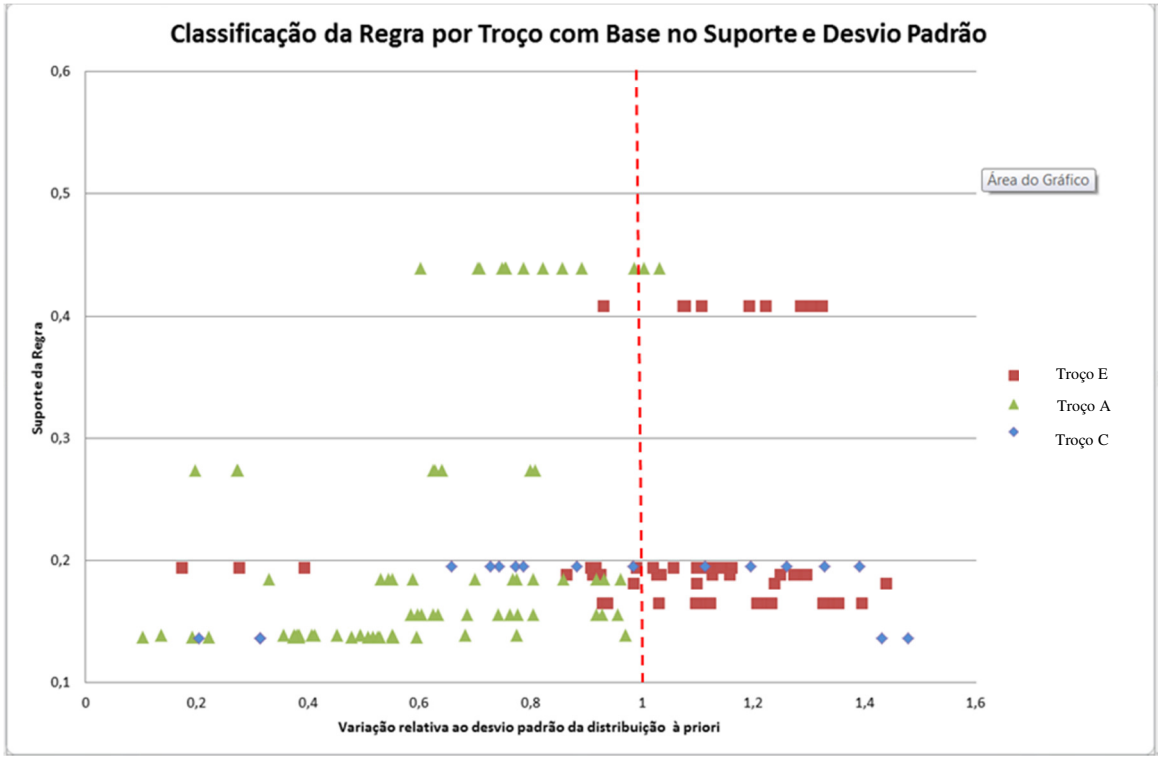


Figura 22: Análise objetiva do interesse das regras

Análise Objetiva

Assim conforme podemos visualizar na Figura 22 o interesse das regras varia no sentido crescente do eixo das ordenadas (suporte da regra) e de acordo com o maior ou menor afastamento relativamente ao valor 1 do eixo das abcissas (assinalada com um traço a tracejado) dado pela variável `X`, que significa o afastamento do desvio padrão da distribuição da regra em presença face ao desvio padrão à priori da distribuição da variável a que esta regra se refere.

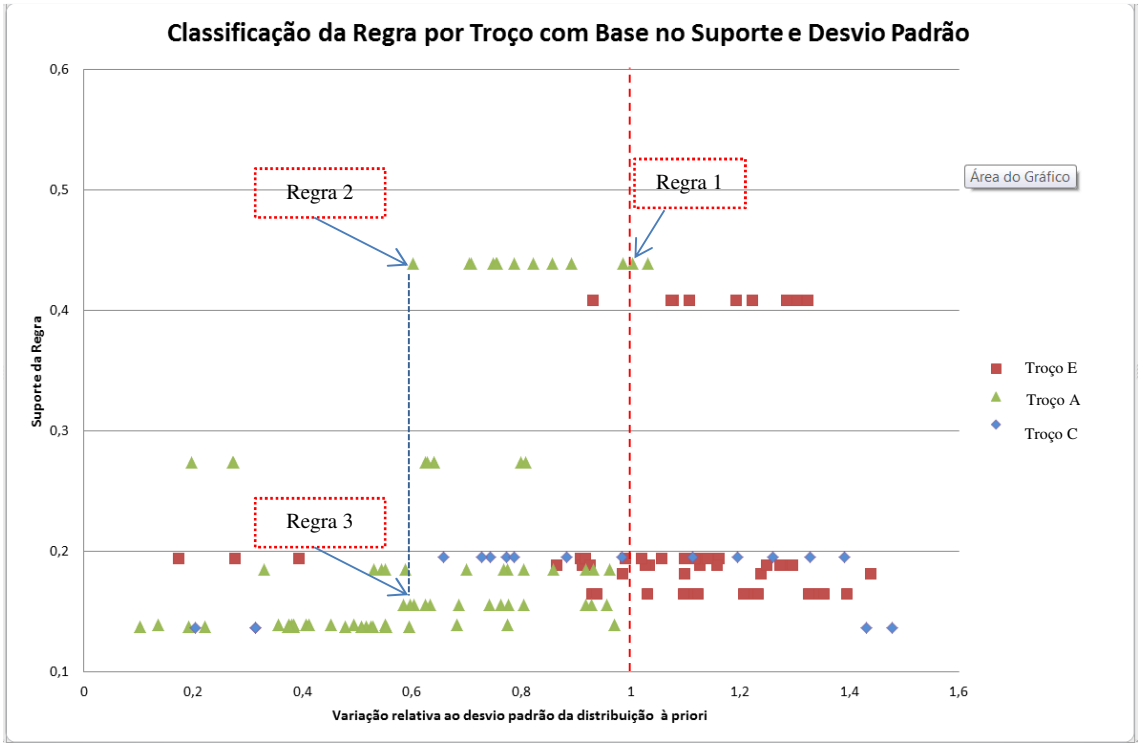


Figura 23: Ordenação do interesse das regras com base em medidas objetivas

Na Figura 23 obtemos a classificação objetiva do interesse das regras 1, 2 e 3. Assim podemos afirmar que a regra 2 é mais importante que a regra 1 e a regra 3, pois:

- Em relação à regra 1, a regra 2 situa-se mais afastada do eixo com origem no valor 1 do eixo das abcissas ou seja com um maior valor de X, que representa o desvio padrão das distribuições à priori de cada uma das regras, e situa-se ao mesmo nível que a regra 1 quando atendemos ao eixo das ordenadas, logo a regra 2 é uma regra com um maior desvio padrão e com um suporte igual à regra 1 assim carateriza-se objetivamente como uma regra mais interessante.
- Em relação à regra 3, a regra 2 situa-se à mesma distância do eixo com origem no valor 1 do eixo das abcissas ou seja um igual valor de X, e situa-se num valor superior ao da regra 3 quando atendemos ao eixo das ordenadas, logo a regra 2 é uma regra com um valor de X igual ao da regra 3 mas com um suporte superior assim carateriza-se objetivamente como uma regra mais interessante.

Análise Subjetiva

A análise subjetiva foi elaborada com base na identificação visual de padrões, subestruturas de dados e o reconhecimento de regras com comportamento "anómalo" face a esses padrões identificados. Estes são aferidos em conjunto com o(s) especialista(s) de domínio.

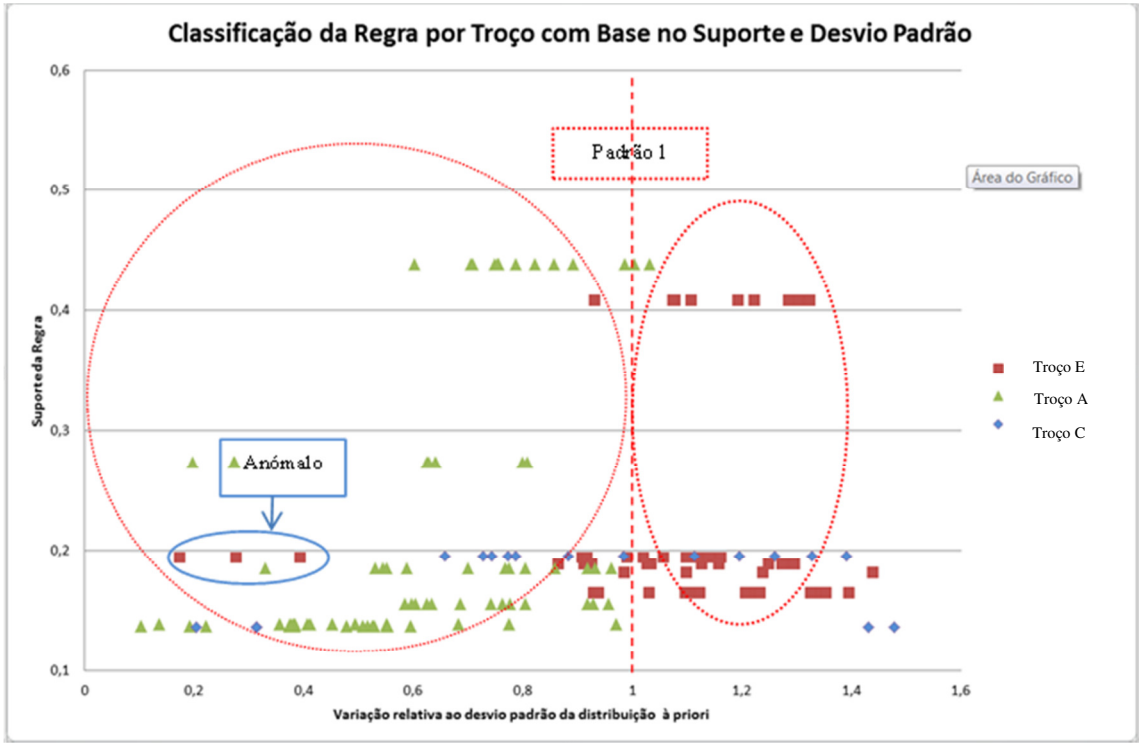


Figura 24: Classificação do interesse das regras com base em análise subjetiva I

Conforme Figura 24 existe um padrão que carateriza as regras do Troço E com um comportamento pior em termos de parâmetros geométricos de via, pois situam-se à direita do referencial com abcissa 1, e as regras do Troço A com um comportamento melhor, pois situam-se à esquerda do referencial. Este padrão foi verificado pelo especialista de domínio confirmando-os. Contudo não foi surpreendente, pois no seu sistema de crenças estas caraterísticas encontravam-se presentes nos troços em apreço.

Contudo verifica-se que um conjunto de regras referentes ao Troço E estão colocadas à esquerda do referencial este é um comportamento anómalo o qual revelou ser surpreendente e portanto interessante, para o especialista de domínio.

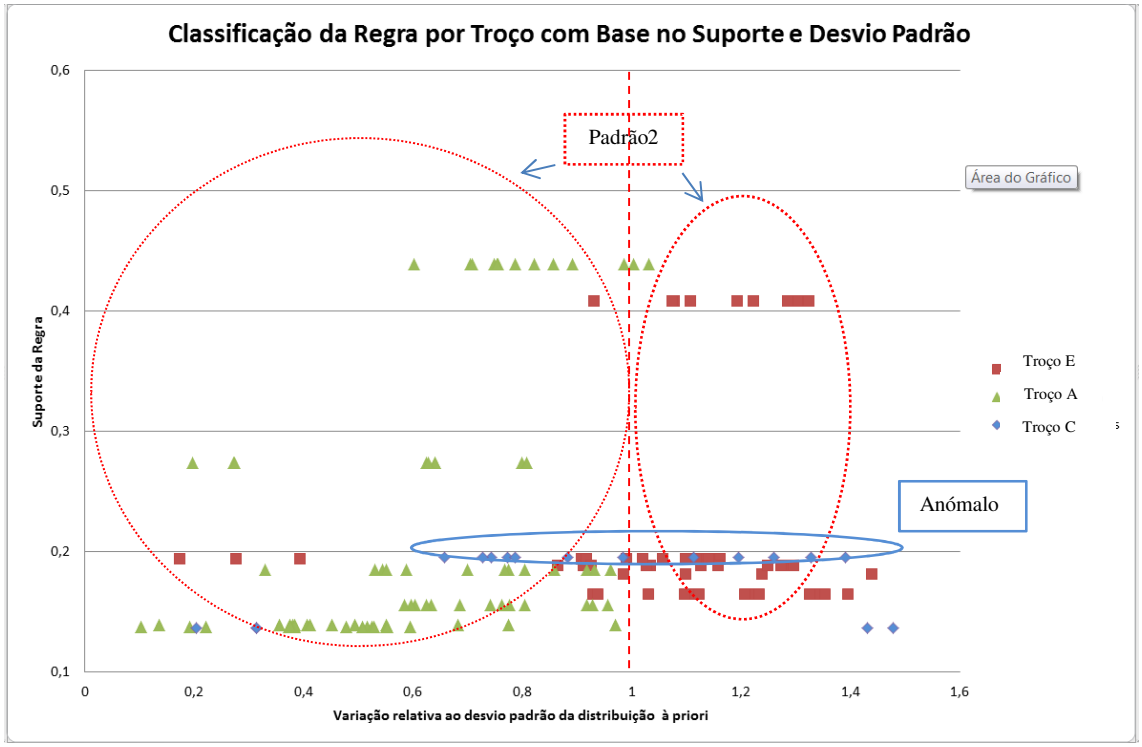


Figura 25: Classificação do interesse das regras com base em análise subjetiva II

Conforme podemos verificar na Figura 25 existe um padrão, as regras de cada troço estão concentradas à esquerda ou à direita da linha com referencial 1 na abcissa, mas existe um terceiro troço que tem um comportamento anómalo face ao padrão instituído, pois as suas regras desenvolvem-se de uma forma transversal ao referencial com abcissa 1. Este padrão é potencialmente interessante.

Consultado o especialista de domínio foi confirmado que a informação presente tinha coerência pois o troço que constituía o padrão anómalo, decompunha-se em uma parte renovada e outra não renovada dando origem a regras à esquerda e à direita do referencial. A informação confirma a crença constituída.

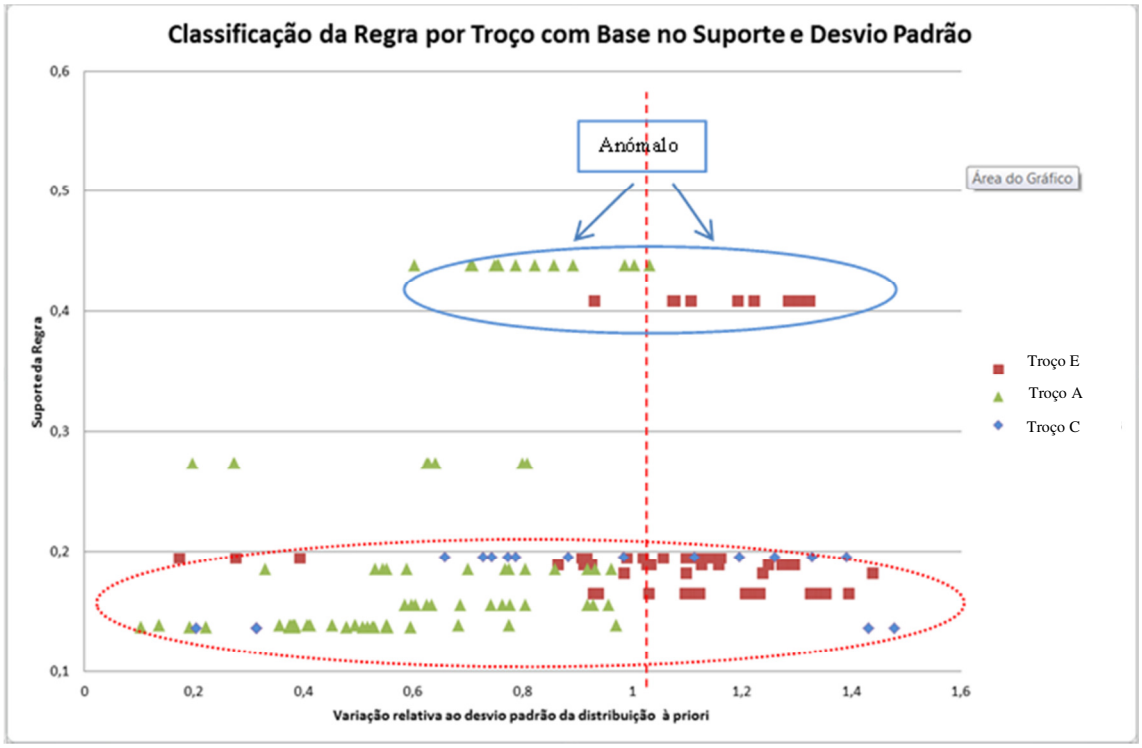


Figura 26: Classificação do interesse das regras com base em análise subjetiva III

Conforme podemos verificar na Figura 26, existe um padrão que corresponde à maior parte das regras de cada troço estarem abaixo do valor 0,2 da coluna das ordenadas, que significa que têm um suporte abaixo dos 20% o que corresponde a cerca de 8 km de via, contudo existem um conjunto de regras com um suporte superior a 16km (≥ 0.4).

Analisadas as regras subjacentes aos dois conjuntos verifica-se que as com suporte superior a 0.4, só contêm um antecedente que é único e caraterizador de cada troço, o que se conclui que estas regras caraterizam os troços em presença, em toda a sua extensão, relativamente a cada um dos parâmetros geométricos em análise e é este facto que determina a anomalia.

O padrão ou seja o conjunto de regras com suporte inferior a 0,2, são caraterizadas por mais de um elemento no antecedente, e embora contenham um elemento antecedente que qualifica a regra como pertencente a um só troço, apresentam outros antecedentes que determinam subconjuntos dentro do referido troço.

Esta perspetiva permite a análise em duas medições, a caraterização de cada troço relativamente a cada um dos parâmetros geométricos de via, e a caracterização de diversos subconjuntos desse troço com caraterísticas próprias, diferenciados entre si.

A perceção destes factos, valida num primeiro momento o conjunto de crenças instituídas, contudo a abordagem, a materialização dos dados, a forma e suporte apresentado foi segundo o especialista de domínio inesperado e disruptivo, pois documenta a complexidade da realidade em observação segundo grandes padrões de comportamento, podendo constituir-se como uma primeira abordagem ao processo de identificação e estudo dos padrões associados à infraestrutura de via.

9. Conclusões

Aproveitando a oportunidade da existência de milhões de dados sobre parâmetros de uma infraestrutura, crendo que no futuro o incremento do registo de dados será exponencial e cada vez mais abrangente, fornecidos pelos processos automatizados de inspeção às infraestruturas, a pertinência da exploração de bases de dados, nomeadamente pelas técnicas de *Data Mining*, será posta na primeira ordem dos processos de aquisição de conhecimento.

Desta forma esta tese visou efetuar uma abordagem de *Data Mining*, a uma área que tradicionalmente procura outras formas de obter conhecimento, esta abordagem foi efetuada através de um conceito relativamente inovador, regras de distribuição à área da gestão/manutenção de infraestruturas nomeadamente infraestrutura ferroviária, com o intuito de obter padrões interessantes.

Assim, enquadrada por uma metodologia que permite a sistematização da abordagem efetuando um alinhamento entre os objetivos do negócio e os objetivos do *Data Mining*, CRISP-DM [Chapman [4]], percorreu-se um percurso desde a compreensão do negócio, aquisição, conhecimento e preparação dos dados, escolha e modelação do algoritmo e avaliação do modelo à luz dos objetivos do negócio, classificando os padrões obtidos de forma objetiva e subjetiva.

O resultado foi a documentação sistematizada de um processo de *Data Mining* aplicada a um caso concreto, que evidencia as dificuldades, técnicas e questões que qualquer abordagem desta área se irá confrontar, culminando na aplicação de um algoritmo, eficiente computacionalmente, que identifica fenómenos estatisticamente diferentes do que é considerado normal à luz dos dados (distribuição à priori) quando em presença de certas caraterísticas da infraestrutura em apreço.

Neste resultado inclui-se uma proposta de apresentação e visualização dos dados modelada pelos parâmetros considerados importantes à luz do negócio, que visa potenciar a completude e a capacidade de interpretabilidade da técnica apresentada,

simplificando uma realidade complexa com o intuito de orientar o especialista de domínio/decisor para a compreensão dos dados e subsequente tomada de decisão.

A questão do "interface" com o utilizador é de absoluta relevância, também no trabalho em apreço, pois permitiu transpor para o utilizador uma nova perspetiva sobre a modelação da realidade da infraestrutura ferroviária, no que concerne aos parâmetros geométricos de via e a sua relação com as caraterísticas/componentes de infraestrutura, de uma forma sucinta, interpretável e significativa, objetivo que julgo atingido.

Trabalho Futuro

Assumindo a presente base de dados, o trabalho imediatamente seguinte será:

- Capacidade de Generalização das Regras

  Aprofundar o trabalho desenvolvido garantindo a obtenção de padrões com a maior capacidade de generalização possível. Desta forma é necessário garantir que a "contaminação" de regras, identificada no ponto 7.3., fruto da existência na base de dados de troços com grandes diferenças no seu comportamento ao nível dos elementos consequentes, não se verifique. Desta forma é necessário revisitar a base de dados inscrevendo um novo atributo, troço, e criar grupos homogéneos de troços no que respeita aos consequentes, seguidamente gerar ficheiros autónomos para cada um dos grupos e reproduzir a análise supra de forma a comparar conjuntos relativamente próximos no que respeita aos parâmetros geométricos de via

- Componente Geográfica

  Aprofundar o trabalho desenvolvido garantindo a obtenção de padrões com uma leitura mais direta com a componente geográfica, revisitar a base de dados inscrevendo um atributo que identifique cada observação por códigos de 200 metros e respetivo troço, de forma a desenvolver uma análise com capacidade de identificação geográfica bem como, mimetizando a análise efetuada

tradicionalmente para esta área de conhecimento, permitir a comparação de resultados.

Cumpre referir que parte do trabalho acima identificado já foi efetuado, nomeadamente no que concerne à capacidade de generalização das regras, já foi inscrito novo atributo, criados grupos homogéneos, Troço A, Troços B,C,D e Troço E e geradas as regras para as competentes variáveis. No que concerne à componente geográfica foi efetuada toda a reconstrução do processo, pois foi necessário ir selecionar o atributo PK aos ficheiros pré integração, para inscrever um atributo que permitisse referenciar os troços de 200 metros. O processo encontra-se no presente, na etapa de discretização faltando discretizar o atributo inclinação.

Trabalho Futuro com Alteração das Atuais Bases de dados

O trabalho a médio e longo prazo resultará da integração dos desenvolvimentos relativos à caraterização da infraestrutura por forma a melhor detalhar os diversos componentes da estrutura de via precisando a relação entre estes e os fenómenos de deterioração dos componentes de via.

A base de dados deverá ser desenvolvida para documentar os momentos de intervenção na estrutura para que, repetindo o processo apresentado neste trabalho, possamos trabalhar com os valores dos parâmetros geométricos de via resultantes da diferença entre duas leituras sendo-lhe associado os valores de MGT que ocorreram entre o espaço de tempo que mediou a sua recolha.

Em paralelo dever-se-á desenvolver a classificação das relações estabelecidas por forma a que, no âmbito de aprendizagem supervisionada seja possível aplicar um sistema de crenças com base numa abordagem *Bayesiana* conforme proposto por [35], aplicando um filtro às regras geradas que permita selecionar as regras através de uma gradação do seu interesse na perspetiva do quanto essa regra é surpreendente.

Apêndice I

Identificação das Bases de Dados

O motivo para a escolha do tema da tese deveu-se ao conhecimento da existência de uma base de dados com origem no processo de leitura automatizado das inspeções que, com periodicidade diversa, semestral no mínimo, percorre a infraestrutura efetuando as leituras dos parâmetros geométricos de via[18]. Desta ação resultam milhões de observações que se destinam a serem classificadas conforme exposto na secção 2.2, para efeitos de monitorização do cumprimento dos parâmetros de adequabilidade da infraestrutura, segundo a norma portuguesa IT.VIA.018 [5], não se encontrando exploradas por nenhuma técnica de extração de conhecimento.

Conforme identificado no anexo I, as bases de dados, relacionadas com o processo de monitorização dos parâmetros de via, dividem-se em dois grandes conjuntos. Um conjunto de dados composto por medições dos atributos que nos indicam os parâmetros geométricos de via e respetivas classificações de alertas denominadas "G_Tables", estas sinalizam a maior ou menor necessidade de correção dos parâmetros, e outro conjunto de dados denominado "G Tables Base" onde se encontram as tabelas que definem a nomenclatura dos parâmetros, as restrições tendentes à integridade das relações, que permite configurar os dois conjunto de dados como uma base de dados relacional.

Acresce referir que as variáveis das bases de dados mencionadas estão expressas em duas escalas de medição, de 25 em 25 cm e de 200 m em 200 m. A escala de 25 cm resulta das leituras efetuadas diretamente pela EM-120 na infraestrutura. A escala de 200 metros resulta da agregação das leituras discretas da escala de 25 cm, contendo em cada observação de 200 metros aproximadamente 800 observações discretas.

Assim temos uma origem de dados em que a fonte são as leituras efetuadas diretamente na infraestrutura proveniente das inspeções efetuadas aos parâmetros de via por veículo

[18] Definição contida no ponto relativo à Terminologia.

de inspeção, cumprindo o normativo europeu EN 13848-5 [(CEN) [6]], denominada de ficheiro G_Medições (ver anexo I), trabalhada subsequentemente em função de:

- Identificação de alertas, através da classificação com base nos parâmetros definidos nos ficheiros G_Tables_Base, das diferentes variáveis resultantes das inspeções constantes do ficheiro G_Medições, permitindo a classificação em diferentes níveis de alertas.

- Identificação de falhas pela agregação espacial de leituras que de forma contínua em espaços não inferiores a 1,25 m (6 observações) e de forma descontínua em espaços não superiores a 4,75 m, situam-se na banda de valores correspondente ao nível de alerta.

- Observações de 200 metros com vista a identificar a estabilidade dos parâmetros pois os alertas e falhas subsequentes são calculados com base nos valores do desvio padrão de cada uma das variáveis em apreço e não no valor da leitura da variável, permitindo aferir a adequabilidade do troço de 200 metros aos parâmetros definidos.

O contexto que carateriza a infraestrutura de via define-se por diferentes tipologias de atributos, associados aos materiais que compõem essa infraestrutura de via, à circulação de comboios que a infraestrutura suporta, ao contexto meteorológico, ao contexto geológico e ao contexto hidrológico.

Efetuada uma pesquisa nas bases de dados disponíveis identificou-se apenas uma base de dados, denominada T_Tables, associada ao contexto da infraestrutura de via sendo esta relativa à circulação de comboios (ver anexo I).

Da base de dados identificada obtém-se a caracterização do tráfego existente em cada linha, segmento, observação, nomeadamente o peso das composições, a velocidade das circulações, comprimento e número.

Apêndice II

Processo de Seleção dos Ficheiros

Processo de seleção do ficheiro G _Medições

Para a seleção do ficheiro G_Medições relativamente ao período temporal de 2012 e para os troços identificados foi necessário decompor o ficheiro que agrega todas as inspeções aos parâmetro geométricos de Via, em ficheiros mais pequenos, para que fosse legível pelo software utilizado, Microsoft Excel, pois a sua dimensão de 7 GB contém milhões de leituras não passíveis de leitura em um só ficheiro no mencionado software:

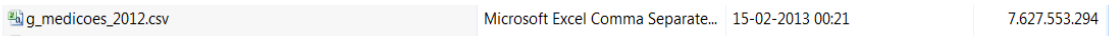


Figura 25: Ficheiro G_Medições_dimensão

Desta forma e recorrendo a uma aplicação denominada "CSV Split" que permite desmultiplicar o ficheiros G_Medições, que contém todas as inspeções de parâmetros geométricos de via relativas ao ano, neste caso 2012, obtemos um conjunto de ficheiros acessíveis à leitura pelo Microsoft Excel. A vicissitude é o número de ficheiros devolvidos pela partição do ficheiro anual, contabilizados para o caso em apreço em 42.

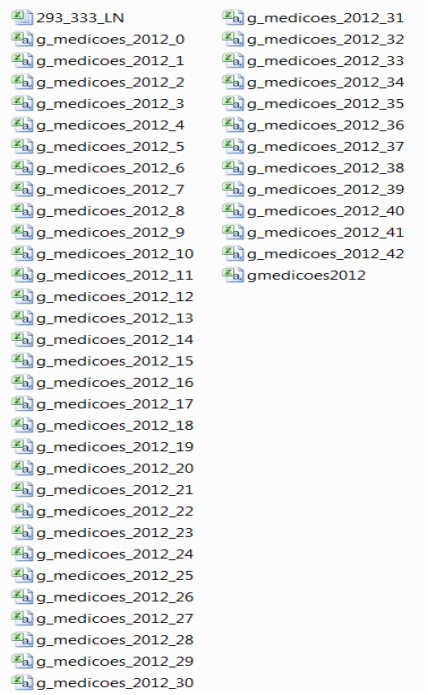


Figura 26: Partição do ficheiro G_Medições

Nestes quarenta e dois ficheiros com um milhão de observações cada é necessário identificar os troços e respetivas datas dos troços a observar que se encontram identificados através de um campo/código inscrito num ficheiro denominado "G_Inspeções_2012" que cataloga as inspeções efetuadas no ano.

ANO	ID	LINHA	COD_HEADER	TROCO	VIA	BITOLA	SENTIDO	DISTANCIA	DATA_LEV
2012	462	8	Troço x	81	VD	1668	D	48698.25	09-07-2012

Figura 27: Túpula do ficheiro G:_Inspeções_2012

O atributo troço constante do ficheiro "G_Inspeções" pode não permitir aceder de uma forma direta ao troço que queremos estudar. Este facto ocorre porque cada inspeção, referenciada pelo atributo "ID" não tem uma correspondência geográfica completa com os troços definidos no ficheiro "G_Inspeções", nem está referenciada dentro dos mesmos, por exemplo através da referência ao ponto quilométrico de início e fim.

Como podemos constatar na Figura 28, o troço x correspondente ao troço 81 na Figura 27, desenvolve-se do Pk 0 ao Pk 300 e conforme Figura 27 só temos indicação que a inspeção desenvolve-se ao longo de 48.698,25 metros.

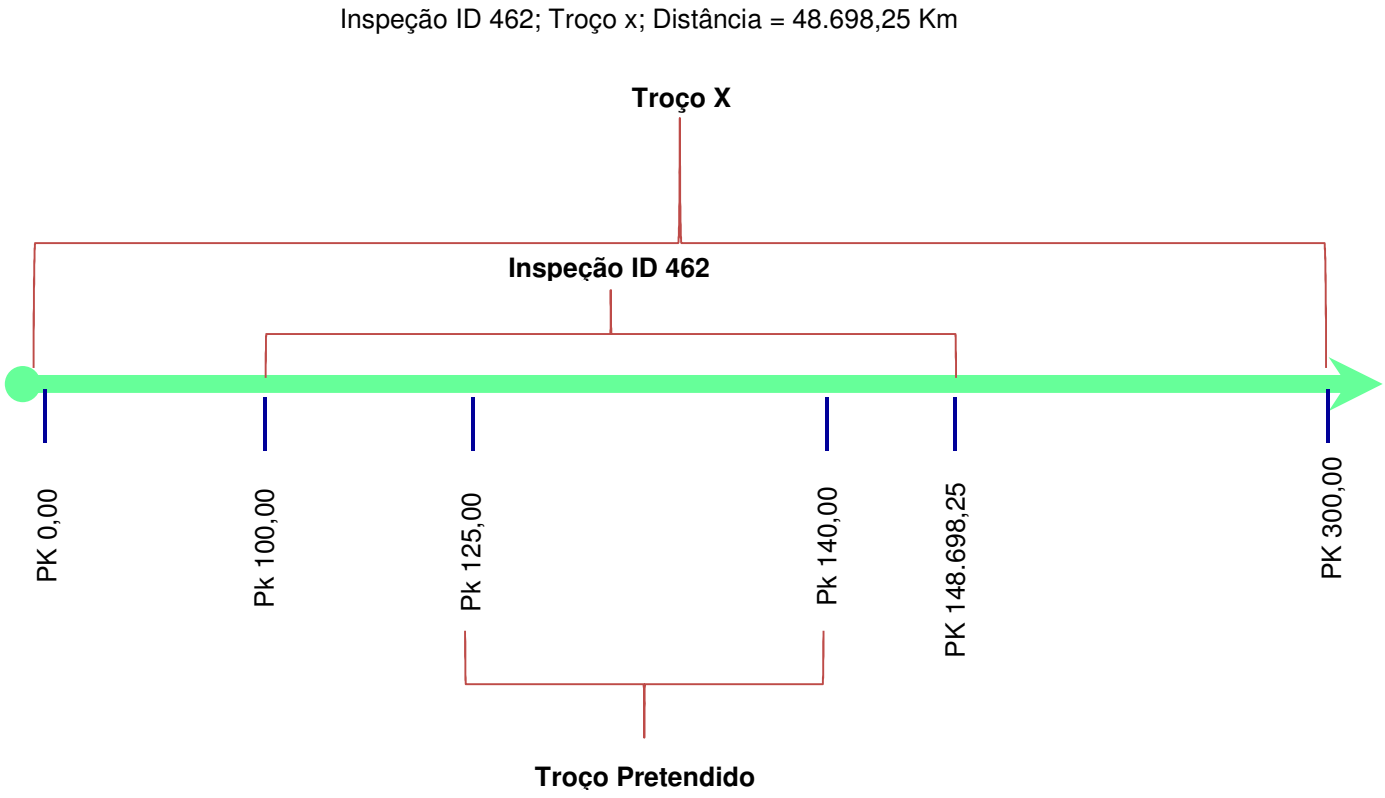


Figura 28: Relação entre inspeção física e ficheiro inspeções

A identificação dos pontos quilométricos da inspeção é efetuada de forma iterativa entre o ficheiro "G_Inspeções" e o ficheiro "G_Medições" atendendo ao atributo "TROÇO" e ao atributo "DISTÂNCIA" do ficheiro "G_Inspeções" e ao atributo "PK"do ficheiro G_Medições. Desta forma a identificação do troço que será objeto de estudo, troço pretendido, é feita por aproximações sucessivas da seguinte forma:

1-Identificação do Troço segundo a nomenclatura do ficheiro "G_Inspeções"

2-Após selecionado no ficheiro "G_Inspeções" o código referente à inspeção que representa o troço e o período requerido, é necessário identificar este código no conjunto dos quarenta e dois ficheiros que compõem o ficheiro "G_Medições".

g_medicoes_2012_0															
g_medicoes_2012_1	499	503	509	536	558	564	581	583	585	591	597	599			
g_medicoes_2012_2	599	602	612	618	619	633									
g_medicoes_2012_3	633	666	667	683	713	714	717	724	728	745	754	755			
g_medicoes_2012_4	755	808	810									955			
g_medicoes_2012_5	955	1089	1105	1114	1118										
g_medicoes_2012_6	366	375	390	445	447	453	470	471	472	1118	1121	1122	1124	1130	1157
g_medicoes_2012_7	472	490	520	554	557	560	567	569	613	621	624	628	630		
g_medicoes_2012_8	628	630	634	637	647	651	652	664	674	687	692	697			
g_medicoes_2012_9	692	735	750	775	782	784	802	804	828	836	845	870	925	938	
g_medicoes_2012_10	342	394	399	406	414	439	446	448	459	938	995	1081	1129	1134	1153
g_medicoes_2012_11	459	469	475	482	483	484	485	500	511	519	524				
g_medicoes_2012_12	511	524	530	545	546	561	566	573	598						
g_medicoes_2012_13	566	600													679
g_medicoes_2012_14	668	675	682	689	691	723	725	730	749	752	753				
g_medicoes_2012_15	753	773	790	805	812	817	858	886	903	934	1031	1053			
g_medicoes_2012_16	353	389	437	442	473	476	480	1053	1096						
g_medicoes_2012_17	480	496	498												
g_medicoes_2012_18	498	507	529	539	543	556	562	574	584	587	608	614	626	631	632
g_medicoes_2012_19	631														690
g_medicoes_2012_20	678	709													769
g_medicoes_2012_21	767														851
g_medicoes_2012_22	832	851	863	887	890	896									
g_medicoes_2012_23	349	368	380	386	396	896	1005	1018	1024	1061	1076	1104	1138	1139	
g_medicoes_2012_24	386	396	427	441	452	461	488	506	513	516	531				
g_medicoes_2012_25	531	571	578	607	611	627	638								
g_medicoes_2012_26	638	653	700	705	707	710	712	726	741						
g_medicoes_2012_27	707	712	715	719	726	741	743	761	762	795	801	829	841	847	848
g_medicoes_2012_28	345	398	404	409	428	440	874	993	1049	1110	1132	1133	1147	1154	1156
g_medicoes_2012_29	428	477	478	479	518	522	547	580	590	606	629	639	645		
g_medicoes_2012_30	645														853
g_medicoes_2012_31	344	392	405	417	436	444	449	487	796	822	837	853	898	997	1012
g_medicoes_2012_32	487	495	527	544	555	570	575	576	577	588	589	593	595	609	616
g_medicoes_2012_33	609	622	676	680	688	708	720	731	740	744	763	764			
g_medicoes_2012_34	763	764	765	776	779	786	806	807	821	825	839				
g_medicoes_2012_35	821	843	846	857	861	891	894	911	1004	1017	1020	1022	1025	1029	1039

Figura 29: Mapeamento do "ID"_Inspeção no G_Medição

Como podemos constatar pela Figura 29 estes códigos de inspeção não se apresentam por ordem cronológica ou sequencial inviabilizando uma pesquisa mais célere.

3º Por último é necessário confrontar os pontos quilométricos (PK) da inspeção selecionada com os PK do troço pretendido (ver Figura 28), pois na própria inspeção pode existir mais que um troço ou apenas segmentos de linha, por forma a retirarmos os respetivos valores dos parâmetros geométricos de via.

Seleção do Ficheiro Circulação

Os ficheiros relativos às variáveis afetas à circulação de comboios, foram obtidos por *query* à base de dados e fornecido ao autor da tese. Contudo importa referir que a qualidade da informação poderia ser melhorada porquanto no processo de obtenção desta variáveis teve o autor desta tese contacto com outra base de dados que continha a velocidade comercial dos comboios e não a velocidade máxima, variável fornecida e que acabou por incorporar o estudo. Contudo a primeira pela diversidade de valores que

apresentava, revela-se uma variável mais interessante do ponto de vista da extração de conhecimento.

Seleção do Ficheiro Diagramas

O ficheiro Diagrama é criado através da transposição manual de valores constantes em suportes físicos, para suporte informático, tendo sido criado manualmente envolveu previamente a seleção de variáveis, melhor detalhado na subsecção 3.2.1. .

Apêndice III

Seleção de Valores de Referência para as Distribuições

Comprimento Total

Conforme podemos constatar na Figura 30, em todos os troços apresentados a distribuição dos valores da variável "Comprimento Total" revela que a mediana, ou percentil 50% (assinalada pela linha horizontal) situa-se descentrada relativamente aos extremos da figura apresentada, o que significa que poderá existir distorção desta medida relativamente à média.

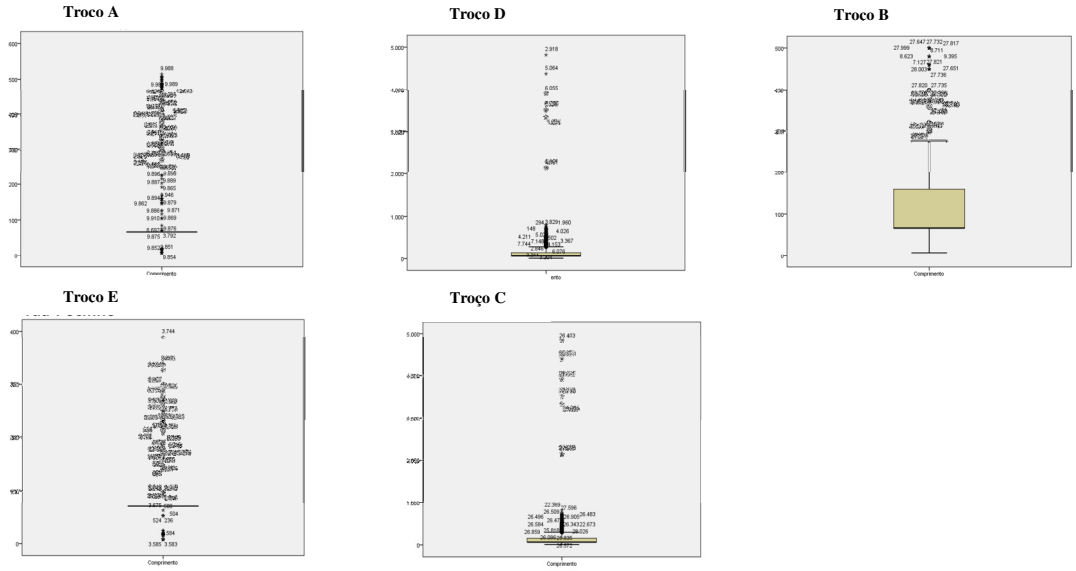


Figura 30: Box Plot atributo "Comprimento Total" por troço

Assim socorrendo-nos das métricas contidas na Figura 31 podemos constatar que para todos os troços a média tem um valor superior entre 7% a 13% quando comparada com a média aparada a 5%, ou seja retirando as 5% observações mais baixas e mais altas e que esta encontra-se mais elevada entre 4% e 46% relativamente à mediana, medida de localização central mais resistente que a média, o que revela que a média encontra-se inflacionada, pelo que não deve ser utilizada como medida de localização central.

Troço A

			Statistic
Comprimento	Mean		84,132
	95% Confidence Interval for Mean	Lower Bound	83,150
		Upper Bound	85,114
	5% Trimmed Mean		77,230
	Median		66,800
	Variance		2522,193
	Std. Deviation		50,2214
	Minimum		7,9
	Maximum		514,6
	Range		506,7
	Interquartile Range		0,0
	Skewness		4,197
	Kurtosis		23,328

Troço D

			Statistic
Comprimento	Mean		124,135
	95% Confidence Interval for Mean	Lower Bound	122,889
		Upper Bound	125,381
	5% Trimmed Mean		110,458
	Median		66,800
	Variance		12426,104
	Std. Deviation		111,4724
	Minimum		7,9
	Maximum		4820,6
	Range		4812,7
	Interquartile Range		92,1
	Skewness		8,871
	Kurtosis		270,721

Troço B

			Statistic
Comprimento	Mean		122,050
	95% Confidence Interval for Mean	Lower Bound	120,901
		Upper Bound	123,200
	5% Trimmed Mean		107,180
	Median		66,800
	Variance		9635,031
	Std. Deviation		98,1582
	Minimum		7,9
	Maximum		500,0
	Range		492,1
	Interquartile Range		92,1
	Skewness		2,195
	Kurtosis		4,542

Troço E

			Statistic
Comprimento	Mean		73,32
	95% Confidence Interval for Mean	Lower Bound	72,01
		Upper Bound	74,62
	5% Trimmed Mean		68,28
	Median		70,00
	Variance		1712,726
	Std. Deviation		41,385
	Minimum		8
	Maximum		389
	Range		381
	Interquartile Range		0
	Skewness		3,288
	Kurtosis		14,270

Troço C

			Statistic
Comprimento	Mean		116,094
	95% Confidence Interval for Mean	Lower Bound	114,852
		Upper Bound	117,336
	5% Trimmed Mean		101,116
	Median		66,800
	Variance		11256,129
	Std. Deviation		106,0949
	Minimum		7,9
	Maximum		4820,6
	Range		4812,7
	Interquartile Range		92,1
	Skewness		11,083
	Kurtosis		364,785

Figura 31: Estatística descritiva atributo "Comprimento Total" por troço

Desta forma conclui-se que o valor médio não é valor de referência para caraterizar as distribuições em apreço, pois encontra-se, ainda que não de uma forma homogénea nos cinco troços, inflacionado pelo peso dos *outliers* existentes na distribuição, como demonstra a análise efetuada às medidas de localização não de tendência central e medidas de dispersão seguidamente analisadas com base na visualização das Figuras 31 e 32.

Como podemos observar na Figura 32 o valor em todos os troços é idêntico até ao 2º quartil, que é correspondente à mediana, o que corresponde a afirmar que 50% das observações são idênticas em todas as distribuições associadas aos diversos troços. O mesmo já não se passa no terceiro quartil cujo comportamento das diferentes distribuições associadas aos diferentes troços é diverso.

Comprimento

	Percentiles						
	5	10	25	50	75	90	95
Troço A	66,800	66,800	66,800	66,800	66,800	158,900	158,900
Troço C	66,800	66,800	66,800	66,800	158,900	216,500	317,700
Troço D	66,800	66,800	66,800	66,800	158,900	250,000	344,600
Troço B	66,800	66,800	66,800	66,800	158,900	230,000	370,000
Troço E	7,90	70,00	70,00	70,00	70,00	70,00	96,00

Figura 32: Medidas de localização não de tendência central atributo "Comprimento Total" por troço

Atendendo à amplitude da série como medida de dispersão para cálculo da variabilidade dos dados e sendo esta calculada como a diferença entre o valor mínimo e máximo da série, atendendo aos valores inscritos na Figura 31 encontramos para os troços em questão as seguintes amplitudes:

Troço	Amplitude
Troço A	506
Troço E	381
Troço C	4813
Troço D	4813
Troço B	492

Tabela 43: Amplitude atributo "Comprimento Total" por troço

Uma medida mais resistente à presença de observações atípicas é a amplitude interquartil,

$$AIQ = Q3 - Q1$$

Expressão (AIII.1): Cálculo da amplitude interquartil

que para os casos em apreço é respetivamente:

Troço	Amplitude AIQ
Troço A	0
Troço E	0
Troço C	92
Troço D	92
Troço B	92

Tabela 44: Amplitude interquartil atributo "Comprimento Total" por troço

Pelo que se conclui que 50% dos elementos centrais nas séries estão contidos num intervalo nulo ou de 92 metros, que nas séries relativas aos Troços A e E o valor mantém-se constante até ao percentil 90, e que em todas as séries a amplitude de 90% dos valores (até ao percentil 90) é significativamente inferior à amplitude global da série, conforme podemos observar na Tabela 45.

	Amplitude	
	Percentil (5-90)	Global
Troço A	92	506
Troço E	62	381
Troço C	150	4813
Troço D	183	4813
Troço B	163	492

Tabela 45: Amplitude interquartil vs amplitude global atributo "Comprimento Total" por troço

Desta forma, atendendo ao comportamento do valor médio face ao valor da média aparada, ou face ao valor da mediana, compreendido que este fenómeno de desvirtuamento carateriza-se pelo facto de que apesar da amplitude da série revelar homogeneidade de valores na maior parte da distribuição e em todos os troços são a presença de *outliers* que desvirtuam a média como medida de localização central, foi escolhida a mediana, que coincide com o valor modal para as séries em apreço, conforme Tabela 46, como valor de referência para caraterizar cada uma das séries em apreço para a variável "Comprimento Total".

Comprimento	Mediana	Valor Modal	Frequência
Troço A	66,8	66,8	84%
Troço D	66,8	66,8	60%
Troço E	70	70	86%
Troço C	66,8	66,8	65%
Troço B	66,8	66,8	64%

Tabela 46: Mediana e valor modal atributo "Comprimento Total" por troço

Velocidade Máxima

Relativamente à variável "Velocidade Máxima", recorro novamente à técnica visual de análise descritiva *Boxplot,* para constatar que a mediana (ou percentil 50) para todos os troços situa-se aproximadamente no meio dos extremos o que significa que não existirá grande distorção desta medida relativamente à média.

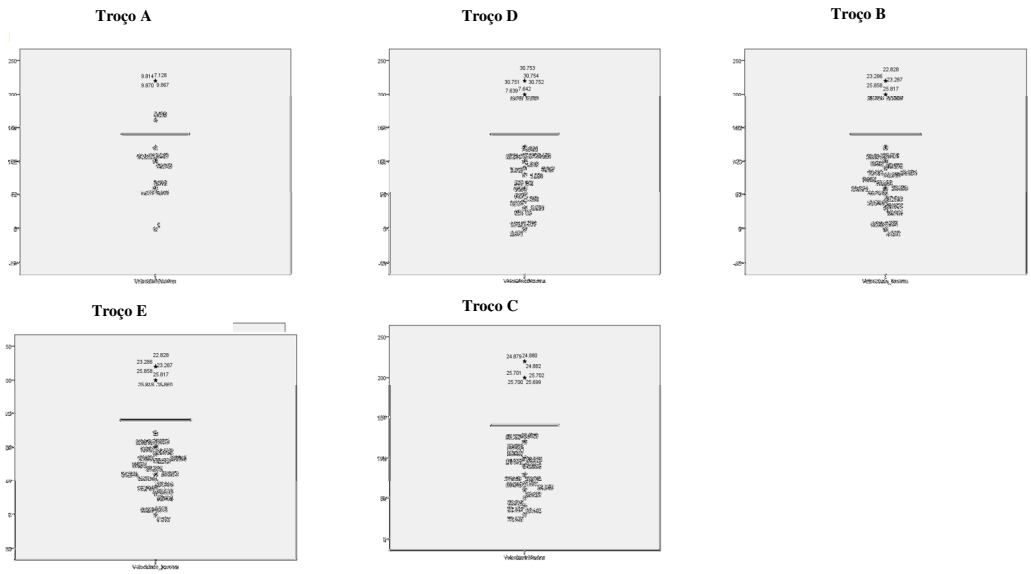


Figura 33: Box Plot atributo "Velocidade Máxima" por troço

Socorrendo-nos das métricas contidas na Figura 34 podemos constatar que para todos os troços, com exceção do E, a média tem um valor superior entre 0,5% a 2% quando comparada com a média aparada a 5%, e que a mediana, medida de localização central mais resistente que a média (citar autor), encontra-se entre 4% e 7,5% abaixo do valor da média, podendo-se considerar também próximo desta, pelo que se confirma a

inferência acima exposta efetuada através da visualização do *Boxplot*, podendo aceitar que a média, no caso desta variável, é uma correta medida de localização central, pois não está influenciada por observações com valores extremos/outliers.

Troço A

			Statistic
VelocidadeMaxima	Mean		149,64
	95% Confidence Interval for Mean	Lower Bound	149,08
		Upper Bound	150,20
	5% Trimmed Mean		147,47
	Median		140,00
	Variance		820,601
	Std. Deviation		28,646
	Minimum		-1
	Maximum		220
	Range		221
	Interquartile Range		0
	Skewness		1,728
	Kurtosis		2,484

Troço D

			Statistic
VelocidadeMaxima	Mean		145,95
	95% Confidence Interval for Mean	Lower Bound	145,54
		Upper Bound	146,36
	5% Trimmed Mean		144,70
	Median		140,00
	Variance		1349,852
	Std. Deviation		36,740
	Minimum		30
	Maximum		220
	Range		190
	Interquartile Range		0
	Skewness		,750
	Kurtosis		-,027

Troço B

			Statistic
VelocidadeMaxima	Mean		151,33
	95% Confidence Interval for Mean	Lower Bound	150,89
		Upper Bound	151,76
	5% Trimmed Mean		150,56
	Median		140,00
	Variance		1304,809
	Std. Deviation		36,122
	Minimum		-1
	Maximum		220
	Range		221
	Interquartile Range		0
	Skewness		,727
	Kurtosis		-,206

Troço E

			Statistic
VelocidadeMaxima	Mean		113,46
	95% Confidence Interval	Lower Bound	112,86
		Upper	114,06
	5% Trimmed Mean		116,52
	Median		120,00
	Variance		358,335
	Std. Deviation		18,930
	Minimum		-1
	Maximum		120
	Range		121
	Interquartile Range		0
	Skewness		-3,005
	Kurtosis		8,589

Troço C

			Statistic
VelocidadeMaxima	Mean		150,44
	95% Confidence Interval	Lower Bound	150,03
		Upper	150,85
	5% Trimmed Mean		149,66
	Median		140,00
	Variance		1249,226
	Std. Deviation		35,344
	Minimum		30
	Maximum		220
	Range		190
	Interquartile Range		0
	Skewness		,769
	Kurtosis		-,001

Figura 34: Estatística descritiva atributo "Velocidade Máxima" por troço

Relativamente ao Troço E e recorrendo às métricas contidas na Figura 34 podemos constatar que a média tem um valor inferior em 2,7% quando comparada com a média aparada a 5%, e que o valor desta encontra-se 6% abaixo do valor da mediana, contudo e à semelhança da decisão tomada para os outros troços, podemos também aceitar a média como valor de referência para esta variável, resultando apenas a nota em sede de análise de estatística descritiva que de forma contrária aos troços remanescentes o valor da média poderá encontrar-se ligeiramente deflacionada como valor de medida central caraterizadora da distribuição.

Apêndice IV

Integração de Dados – Problema de Resolução da Identidade

Para a integração do ficheiro G_Medições resultante do processo de seleção descrito no apêndice II com os outros dois ficheiros Circulação e Diagramas de Via é necessário identificar um campo que seja comum a todos e que subsequentemente permita a integração.

ANO	ID_INSP	DIST_O	ID_POS	KM	LOCALIZ	PK	P_SPEED	F_SPEED	SPEED

Figura 35: Mapeamento do "ID"_Inspeção no G_Medição

A Figura 35 identifica os campos pertencentes ao ficheiro G_Medições que não são parâmetros geométricos de via, são elementos caraterizadores da recolha de dados. Entre estes, os candidatos a um campo comum que possibilite a integração entre os ficheiros Circulação e Diagrama de Via, só pode ser um campo que expresse uma referência geográfica, pois definido o período temporal como o ano de 2012, só temos que garantir a integração espacial, ou seja que os dados constantes do ficheiro G_Medições digam respeito ao mesmo ponto geográfico que os dados constantes do ficheiro Circulação e Diagrama de Via.

Desta forma dos atributos acima expostos os candidatos a referencial para a integração são o atributo "KM", o atributo "LOCALIZ" e o atributo "PK" que é o acrónimo de ponto quilométrico. Conforme esclarecimentos efetuados pelo especialista de domínio, o atributo "KM" é introduzido à mão pelo operador do veículo de recolha dos dados sempre que este entenda efetuá-lo sendo calculado subsequentemente através do atributo "LOCALIZ", o atributo "LOCALIZ" é calculado pelo odómetro[19] com base no "KM" definido pelo operador, o campo "PK" é a junção destes dois para se obter um referencial geográfico completo.

[19] Equipamento que mede a distância percorrida

Desta forma foi escolhido como referencial o campo "PK" pois identificava ao centímetro a observação em causa o que permitia a integração com os ficheiros remanescentes.

Após ter escolhido o campo "PK" como referencial de integração constatei que o mesmo apresentava inconsistências, nomeadamente sendo a leitura no sentido descendente, o "PK" evoluía em sentido inverso, como também existiam observações com o mesmo valor no campo "PK".

ANO	ID_INSP	DIST_O	DIST_O_Trab	ID_POS	KM	LOCALIZ	PK	Pk vs Pk Trabalhado	PKTrabalhad	PKTrabalh	Confirmaç	P_SPEED	F_SPEED	SPEED	NIVLE	NIVLD	NIVLED1	NIVLDD1	NIVLED2	NIVLDD2	ALIN
2012	373	.00	0	0	38	200.00	3.820.000	0	3.820.000	38.200	38.200	0	120	.00							-13.1
2012	373	.25	0,25	0	38	200.00	3.820.000	-25	3.820.025	38.200,3	38.200	0	120	.00							-13.0
2012	373	.50	0,5	0	38	200.25	3.820.025	-25	3.820.050	38.200,5	38.200	0	120	.00							-12.7
2012	373	.75	0,75	0	38	200.50	3.820.050	-25	3.820.075	38.200,8	38.200	0	120	.00							-12.4
2012	373	1.00	1	1	38	200.75	3.820.075	-25	3.820.100	38.201,0	38.200	0	120	.00							-12.0

Figura 36: Exemplo de inconsistência do campo PK relativamente ao campo DIST_O (observações com o mesmo valor no campo PK)

ANO	ID_INSP	DIST_O	DIST_O_Trab	ID_POS	KM	LOCALIZ	PK	Pk vs Pk Trabalhado	PKTrabalhac	PKTrabalh	Confirmaç	P_SPEED	F_SPEED	SPEED	NIVLE	NIVLD	NIVLED1	NIVLDD1	NIVLED2	NIVLDD2	ALIN
2012	373	6807.50	6807,5	6807	44	995.00	4.499.500	-1.250	4.500.750	45.007,5	38.200	0	130	105.23	-.82	-.55	-.78	-.35	-3.32	-.94	16.6
2012	373	6807.75	6807,75	6807	44	995.25	4.499.525	-1.250	4.500.775	45.007,8	38.200	0	130	105.23	-.78	-.62	-.66	-.39	-3.24	-.82	16.5
2012	373	6808.00	6808	6808	44	995.50	4.499.550	-1.250	4.500.800	45.008,0	38.200	0	130	105.23	-.78	-.74	-.55	-.47	-3.12	-.74	16.5
2012	373	6808.25	6808,25	6808	44	995.75	4.499.575	-1.250	4.500.825	45.008,3	38.200	0	130	105.23	-.70	-.82	-.51	-.55	-3.01	-.62	16.4
2012	373	6808.50	6808,5	6808	44	996.00	4.499.600	-1.250	4.500.850	45.008,5	38.200	0	130	105.23	-.82	-1.02	-.55	-.70	-2.89	-.55	16.5
2012	373	6808.75	6808,75	6808	44	996.25	4.499.625	-1.250	4.500.875	45.008,8	38.200	0	130	105.23	-.90	-1.13	-.62	-.82	-2.77	-.43	16.7
2012	373	6809.00	6809	6809	44	996.50	4.499.650	-1.250	4.500.900	45.009,0	38.200	0	130	105.23	-.94	-1.25	-.66	-.90	-2.66	-.35	16.8
2012	373	6809.25	6809,25	6809	44	996.75	4.499.675	-1.250	4.500.925	45.009,3	38.200	0	130	105.23	-.98	-1.29	-.62	-.94	-2.50	-.23	17.0
2012	373	6809.50	6809,5	6809	45	.00	4.500.000	-950	4.500.950	45.009,5	38.200	0	130	105.23	-.78	-1.21	-.51	-.86	-2.38	-.12	17.0
2012	373	6809.75	6809,75	6809	45	.25	4.500.025	-950	4.500.975	45.009,8	38.200	0	130	105.23	-.66	-1.13	-.31	-.70	-2.23	-.04	16.9
2012	373	6810.00	6810	6810	45	.50	4.500.050	-950	4.501.000	45.010,0	38.200	0	130	105.23	-.51	-1.05	-.12	-.62	-2.11	.04	16.8

Figura 37: Exemplo de inconsistência do campo PK relativamente ao campo DIST_O (o campo PK evolui em uma distância superior)

Tendo consultado o especialista do domínio sobre este facto este acrescentou que o referencial de qualidade em termos de distâncias seria o atributo "DIST_O"[20] pois este é medido diretamente pelo sensor das rodas, contudo expressando este atributo uma distância relativamente à origem da inspeção não constituía um referencial de localização. Desta forma e atendendo que o único atributo que permitia ser um referencial completo "PK" apresentava inconsistências, foi criado um novo atributo

[20] Foi necessário converter o atributo "DIST_O" de formato texto para formato numério, criando mais um atributo, pois só em formato numérico foi possível hierarquizar através do atributo "DIST_O" o ficheiro desde o ponto incial de inspeção até ao final.

denominado “PKTrabalhado” que é função, para todas as observações, da soma do atributo “DIST_O” ao valor do atributo “PK” para a observação onde o atributo “DIST_O” tem valor zero, ou seja onde segundo o atributo “DIST_O” se iniciou a inspeção.

Foi este novo atributo “PKTrabalhado” que passou a ser o referencial para a integração dos ficheiros nomeadamente entre os ficheiros “G_Medições” e “Diagramas de Via”.

Bibliografia

[1] Ministério da Economia e do Emprego, M. *Plano Estratégico dos Transportes: Mobilidade Sustentável. Horizonte 2011-2015*. INCM, City, 2011.
[2] Jorge, A. M., Azevedo, P. J. and Pereira, F. *Distribution Rules with Numeric Attributes of Interest*. Springer Berlin Heidelberg, City, 2006.
[3] Agrawal R, I. T., Swami A Mining Association Rules between Sets of Items in Large Databases. *SIGMOD Rec.*, 22, 2 1993), 207-216.
[4] Chapman, J. C. R. K. T. K. T. R. C. S. R. W. P. *CRISP-DM 1.0 - Step-by-step data mining guide*. SPSS, 2000.
[5] REFER *IT.VIA.018*. REFER, City, 2009.
[6] (CEN), E. C. f. S. *European Standard EN 13848-5: Railway applications – Track – Track geometry quality – Part 5: Geometric quality levels.*, City, 2008.
[7] Andrade, A. R. *Prediction and optimization of maintenance and renewal actions related to rail track geometry* IST, UL_Instituto Superior Tecnico, 2014.
[8] UIC *Best practice guide for optimum track geometry durability*. City, 2008.
[9] João Gama, A. C. L., Katti Faceli, Andre Ponce de Leon Carvalho,Márcia Oliveira *Extração de Conhecimento de Dados*. Edições Sílabo, lda, 2012.
[10] Freund, Y., Iyer, R., Schapire, R. E. and Singer, Y. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 42003), 933-969.
[11] López Pita, A., Teixeira, P.F., Casas, C., Ubalde, L. and Robusté, F. Evolution of track geometric quality in high-speed lines: ten years experience of the Madrid-Seville line. *Proc.IMechE, Part F, Journal of Rail and Rapid Transit, vol. 221, No. 1, pp. 147-155.* 2007).
[12] Esveld, C. *Modern Railway Track*, 2001.
[13] Gross, P., Boulanger, A., Arias, M., Waltz, D. L., Long, P. M., Lawson, C., Anderson, R., Koenig, M., Mastrocinque, M. and Fairechio, W. *Predicting electricity distribution feeder failures using machine learning susceptibility analysis*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, City, 2006.
[14] Long, P. M. and Servedio, R. A. *Martingale boosting*. Springer Berlin Heidelberg, City, 2005.
[15] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G. *Learning to rank using gradient descent*. ACM, City, 2005.
[16] Molina, L. F., Resendiz, E., Edwards, J. R., Hart, J. M., Barkan, C. P. and Ahuja, N. *Condition monitoring of railway turnouts and other track components using machine vision*. City, 2011.
[17] Resendiz, E., Hart, J. M. and Ahuja, N. Automated Visual Inspection of Railroad Tracks2013).
[18] Marino, F. and Stella, E. ViSyR: A vision system for real-time infrastructure inspection. *Vision Systems: Applications*2007), 113-144.
[19] SHAH, D. M. Automated Visual Inspection/Detection of Railroad Track2010).
[20] Alippi, C., Casagrande, E., Scotti, F. and Piuri, V. Composite real-time image processing for railways track profile measurement. *Instrumentation and Measurement, IEEE Transactions on*, 49, 3 2000), 559-564.
[21] Xishi, W., Bin, N. and Yinhang, C. *A new microprocessor based approach to an automatic control system for railway safety*. IEEE, City, 1992.
[22] Guclu, A., Yılboga, H., Eker, O. F., Camci, F. and Jennions, I. Prognostics with Autoregressive Moving Average for Railway Turnouts. *Annual Conference of the Prognostics and Health Management Society*2010).
[23] García Márquez, F. P., Schmid, F. and Conde Collado, J. A reliability centered approach to remote condition monitoring. A railway points case study. *Reliability Engineering & System Safety*, 80, 1 2003), 33-40.
[24] Garcia Marquez, F. P., Pedregal Tercero, D. J. and Schmid, F. Unobserved component models applied to the assessment of wear in railway points: A case study. *European journal of operational research*, 176, 3 2007a), 1703-1712.
[25] Roberts, C., Dassanayake, H., Lehrasab, N. and Goodman, C. Distributed quantitative and qualitative fault diagnosis: railway junction case study. *Control Engineering Practice*, 10, 4 2002), 419-429.
[26] García Márquez, F. P. and Schmid, F. A digital filter-based approach to the remote condition monitoring of railway turnouts. *Reliability Engineering & System Safety*, 92, 6 2007b), 830-840.

[27] Atamuradov, V., Camci, F., Baskan, S. and Sevkli, M. *Failure diagnostics for railway point machines using expert systems*. IEEE, City, 2009.
[28] Márquez, F. G., Roberts, C. and Tobias, A. M. Railway point mechanisms: condition monitoring and fault detection. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 224, 1 2010), 35-44.
[29] Camci, F. and Chinnam, R. Process Monitoring, Diagnostics and Prognostics in Machining Processes. *LAP Lambert Academic Publishing: Saarbrücken, Germany*2010a), 978-3838335667.
[30] Camci, F. and Chinnam, R. B. Health-state estimation and prognostics in machining processes. *Automation Science and Engineering, IEEE Transactions on*, 7, 3 2010b), 581-597.
[31] Eker, O. F., Camci, F. and Kumar, U. *Failure Diagnostics on Railway Turnout Systems Using Support Vector Machines*. City, 2010.
[32] Yilboga, H., Eker, O., Güçlü, A. and Camci, F. *Failure prediction on railway turnouts using time delay neural networks*. IEEE, City, 2010.
[33] Fayyad, U. and Irani, K. Multi-interval discretization of continuous-valued attributes for classification learning1993).
[34] Aumann, Y. and Lindell, Y. A statistical theory for quantitative association rules. In *Proceedings of the Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (San Diego, California, USA, 1999). ACM, [insert City of Publication],[insert 1999 of Publication].
[35] Wong, J. A. H. M. A. Algorithm AS 136: A K-Means Clustering Algorithm *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 1 1979), 100-108.
[36] Webb, G. I. Discovering associations with numeric variables. In *Proceedings of the Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (San Francisco, California, 2001). ACM, [insert City of Publication],[insert 2001 of Publication].
[37] Srikant, R. and Agrawal, R. Mining quantitative association rules in large relational tables. *SIGMOD Rec.*, 25, 2 1996), 1-12.
[38] Piateski, G. and Frawley, W. *Knowledge Discovery in Databases*. MIT Press, 1991.
[39] and, R. A. and Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. *VLDB'94, Proceedings of 20th International Conference on Very*
*Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*1994), 487-499.
[40] Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T. Mining optimized association rules for numeric attributes. In *Proceedings of the Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (Montreal, Quebec, Canada, 1996). ACM, [insert City of Publication],[insert 1996 of Publication].
[41] Morimoto, Y., Ishii, H. and Morishita, S. Efficient Construction of Regression Trees with Range and Region Splitting. *Machine Learning*, 45, 3 (2001/12/01 2001), 235-259.
[42] Zhang, Z., Lu, Y. and Zhang, B. *An effective partitioning-combining algorithm for discovering quantitative association rules*. City, 1997.
[43] Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T. Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization. *SIGMOD Rec.*, 25, 2 1996), 13-23.
[44] Yoda, K. Computing Optimized Rectilinear Regions for AssociationRules. *Proc. of KDD-97 : Third Int. Conf. on Knowledge Discovery and Data Mining*(1997 1997), 96-103.
[45] Brin, S., Motwani, R., Ullman, J. D. and Tsur, S. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the Proceedings of the 1997 ACM SIGMOD international conference on Management of data* (Tucson, Arizona, USA, 1997). ACM, [insert City of Publication],[insert 1997 of Publication].
[46] Silberschatz, A. and Tuzhilin, A. *On subjective measures of interestingness in knowledge discovery*. City, 1995.
[47] C. Alchourron, P. G., and D. Makinson On the logic of theory change: Partial meet contraction and revision functions.
bolic Logic. *Journal of Symbolic Logic*(1985 1985), 510-530.
[48] Mcgarry, K. A Survey of Interestingness Measures for Knowledge Discovery. *The Knowledge Engineering Review*, Vol. 00:0, 1–242005).
[49] Ludwig, J. a. L., G. Whats new? using prior models as a measure of novelty
in knowledge discovery. *Proceedings of the 12th IEEE Conference on Tools with Artificial Intelliegnce'*2000), 86-89.

Anexos

Anexo I

Bases de Dados

G-Tables – Conjunto de ficheiros com as leituras de geometria da via

.

- G_Eventos - Ficheiro com os marcos físicos na infraestrutura identificados manualmente pelos operadores da M120, como sejam fim de estação, inicio passagem de nível, início AMV (labels).
- G_ Inspeções –Ficheiro com a identificação, por troço, das inspeções efetuadas à via no ano.
- G_Med_Alert – Ficheiro com os alertas resultantes da aplicação dos parâmetros definidos nos ficheiros g_tables_base, é produzido um alerta por cada leitura que ultrapasse os limites impostos pelos parâmetros definidos.
- G_Med_Fault –Ficheiro com a Identificação dos defeitos com base nos alertas identificados na tabela G_Med_Alert, atendendo ao comprimento mínimo que uma falha (alerta) deve comportar para ser considerado defeito. Transforma a falha em alerta.
- G_Med_200 – Ficheiro com os segmentos de 200 metros com os cálculos de desvio padrão para cada uma das variáveis objeto de medição, cada registo é um segmento de 200m.
- G_Med_200_Class – Ficheiro com as classificações atribuídas aos dados de 200m com base nos ficheiros G-Tables Base
- G_Medições – Ficheiro com as leituras efetuadas de 25 em 25 cm para cada um dos parâmetrosque suportam a análise de geometria de via que constituem a base dos dados deste trabalho de *Data* Mining referentes a todas as inspeções efetuadas no ano n.
- G_Posição – Ficheiro com as coordenadas da posição GPS de cada ponto de medição.

G_Tables_Base – Conjunto de ficheiros com as tabelas de referência que permitem classificar os ficheiros com as leituras de geometria de via**.**

- G_Bitolas – Ficheiro com a identificação e códigos de tipos de bitola.
- G_Classe_velocidade – Ficheiro com a identificação e códigos das diferentes classes de velocidade.
- G_Meta_param – Ficheiro com a identificação e códigos da tolerância de valores de leituras, meta parâmetros, para diferenciar os critérios de aceitação de linhas novas ou renovadas, trabalhos de manutenção e ações de manutenção.

- G_Meta_param_200 Ficheiro com a identificação e códigos da tolerância de valores de leituras, meta parâmetros para 200 metros, para identificar Nível de Qualidade e Alertas.
- G_param_200_med - Ficheiro com a identificação dos códigos para as tabelas de tolerância sobre os desvios padrão num bloco de 200m, as leituras são agregadas em desvios padrão para segmentos de 200m segundos dois critérios os 200m de PK e os 200 m de distância real ao centro dos 200m.
- G_param_med – Ficheiro com a identificação dos códigos para cada um dos parâmetros que suportam a análise de geometria de via.
- G_parâmetros – Ficheiro com a identificação dos códigos que identificam os tipos de parâmetros que suportam a análise de geometria de via.
- G_parâmetros_200 - Ficheiro com a identificação dos códigos que identificam os tipos de parâmetros que suportam a análise de geometria de via em segmentos de 200 m .
- G_status - Ficheiro com a identificação dos códigos que identificam os diversos níveis de alerta.
- g_status_200 - Ficheiro com a identificação dos códigos que identificam os diversos níveis de alerta para segmentos de 200m
- G_tolerancias_param **–** Ficheiro que define os intervalos de tolerância de cada um dos parâmetros e associa os códigos que identificam os diversos níveis de alerta.
- G_tolerancias_param_200 - Ficheiro que define os intervalos de tolerância de cada um dos parâmetros e associa os códigos que identificam os diversos níveis de alerta para segmentos de 200m.

Diagramas de Via - Variáveis documentadas

- Velocidade
  - Máxima
  - Máxima de Planta
- Quilometragem
- Número das Curvas
- Pontos de Tangência
- Diagrama de Curvas - (Visual)
- Caraterísticas
  - Velocidades (Máxima de Planta)
  - Raio
  - Escala

- Insuficiência de Escala
- Disfarce de Escala
- Variação de Insuficiência no Tempo
- Variação de flechas

- Armamento
- Carril
- Travessa

- Perfil Longitudinal
- Inclinação
- Comprimento
- Raio
- Bissetriz

Diagrama de Estações e Pontos Singulares – Localização

- Comprimento

T_Tables – Conjunto de ficheiros com informação sobre circulação.

- T_Causas do Atraso Detalhes - Ficheiro com a identificação dos códigos que identificam as diversas causas de atrasos.
- T_Causas do Atraso Grupo - Ficheiro com a identificação dos códigos que identificam os diversos grupos a imputar os atrasos.
- T_Causas do Atraso Sub-Grupo - Ficheiro com a identificação dos códigos que identificam os diversos sub-grupos a imputar os atrasos.
- T_Dependências - Ficheiro com a identificação dos códigos que identificam as dependências no que concerne à circulação.
- T_Histórico Infraestrutura - Ficheiro com a identificação dos códigos que identificam diferentes períodos temporais da infraestrutura.
- T_Localiza - Ficheiro que relaciona as dependências registadas no ficheiro T_Depenências com a identificação dos códigos dos diferentes períodos temporais da infraestrutura constante do ficheiro T_Histórico_Infraestrutura atribuindo a localização às mencionadas dependências através da identificação da linha e do ponto quilométrico.
- T_Material_Motor - Ficheiro com a identificação dos códigos que identificam a tipologia da unidade motora bem como as suas caraterísticas.

- T_Movimento Comboios - Ficheiro com a identificação dos códigos que identificam os movimentos dos comboios.
- T_Movimento Comboios_Atraso - Ficheiro com a identificação dos códigos que identificam as causas do atraso dos movimentos dos comboios.
- T_Movimento Comboios_Composição - Ficheiro que relaciona o código do movimento comboio do ficheiro T_Movimento Comboios, com o código do material motor do ficheiro T_Material_Motor, com o código das dependências do ficheiro T_Dependências, com as variáveis carga rebocada, comprimento rebocado, velocidade.
- T_Movimento_Comboios_Itenerário - Ficheiro que relaciona o código dos movimentos comboio constantes do ficheiro T_Movimento Comboios com os códigos das dependências constantes do ficheiro T_Dependências com os códigos do tipo de Movimento constante do Ficheiro T_Tipo_Movimento_Comboio, estabelecendo um itinerário para o comboio.
- T_Operadores – Ficheiro que identifica os códigos dos operadores passíveis de circularem.
- T_Responsáveis_do_Atraso – Ficheiro que identifica os códigos dos operadores responsáveis pelos atrasos.
- T_Tipo_Movimento_Comboios – Ficheiro que identifica os códigos do tipo de movimento dos comboios.
- T_Tipo_Serviços – Ficheiro que identifica os códigos do tipo de serviços
- T_Vias – Ficheiro que identifica os códigos da linha, relacionando com o código das dependências constantes do ficheiro T_Dependências, atribuindo um código de via que identifica o sentido ascendente ou descendente.

Anexo II

Ficheiros Selecionados

G_Medições (Inicial)
ANO
ID_INSP
DIST_O
ID_POS
KM
LOCALIZ
PK
P_SPEED
F_SPEED
SPEED
NIVLE
NIVLD
NIVLED1
NIVLDD1
NIVLED2
NIVLDD2
ALINE
ALIND
ALINED1
ALINDD1
ALINESQR
ALINDIRR
NIVTRANS
NIVTRANR
EMPENO3M
EMPREL3M
EMPENO9M
EMPREL9M
BITOLA
BITMED
GRAD
GRADMED
CURVA

Circulação (Inicial)
NumeroComboio1
NumeroComboio2
DataRealizacao
OperadorId
Operador
RegimeFrequenciaMnemonica
TipoServicoId
TipoServico
DataInicioCirculacao
DataFimCirculacao
DocumentoHorario
DependenciaOrigemId
DependenciaOrigemDescricao
DependenciaDestinoId
DependenciaDestinoDescricao
HoraPartida
HoraChegada
IsSuprimidoTotal
IsSuprimidoParcial
CargaRebocada
ComprimentoRebocado
IdMaterialMotor1
VelocidadeMaxima
Sentido
Comprimento

Troço (Inicial)
PK
Inclinação
Curva de Concordância
Carril
Carril II
Travessa
Vel. Máxima de Planta

Anexo III

Ficheiro G_Medições

G_Medições (Inicial)	Definição do Atributo
ANO	Ano a que dizem respeito as observações
ID_INSP	Código da identificação da inspeção
DIST_O	Distância à origem
ID_POS	Identificação da posição da observação (manual)
KM	Quilómetro a que dizem respeito as leituras
LOCALIZ	Metros a que dizem respeito as leituras
PK	Agrega os valores dos atributos KM + Localização
P_SPEED	Vel máxima do troço segundo Tabela de Velocidades Máximas
F_SPEED	Vel instantânea da máquina de inspeção
SPEED	Velocidade a que a máquina inspecionou
NIVLE	Nivelamento Longitudinal da Fila Esquerda
NIVLD	Nivelamento Longitudinal da Fila Direita
NIVLED1	Nivelamento Longitudinal da fila esquerda com um comprimento de onda entre 1m e 25m.
NIVLDD1	Nivelamento Longitudinal da fila esquerda com um comprimento de onda entre 1m e 25m.
NIVLED2	Nivelamento Longitudinal da fila esquerda com um comprimento de onda entre 25m e 70 metros.
NIVLDD2	Nivelamento Longitudinal da fila direita com um comprimento de onda entre 25m e 70 metros.
ALINE	Alinhamento da fila esquerda
ALIND	Alinhamento da fila direita
ALINED1	Alinhamento da fila esquerda com base em cordas de 10 metros.
ALINDD1	Alinhamento da fila direita com base em cordas de 10 metros.
ALINESQR	Alinhamento relativo da fila esquerda
ALINDIRR	Alinhamento relativo da fila direita
NIVTRANS	Nivelamento Transversal (de uma fila de carril relativamente à outra)
NIVTRANR	Nivelamento Transversal (relativamente ao plano horizontal de referência)
EMPENO3M	Empeno a 3m
EMPREL3M	Empeno relativo a 3m
EMPENO9M	Empeno a 9m
EMPREL9M	Empeno relativo a 9m
BITOLA	Bitola
BITMED	Bitola Média
GRAD	Gradiente
GRADMED	Gradiente Médio
CURVA	Curva

Anexo IV

Preparação dos Dados (Construção e Seleção) – Ficheiro Circulação

Circulação (Inicial)
NumeroComboio1
NumeroComboio2
DataRealizacao
OperadorId
Operador
RegimeFrequenciaMnemonica
TipoServicoId
TipoServico
DataInicioCirculacao
DataFimCirculacao
DocumentoHorario
DependenciaOrigemId
DependenciaOrigemDescricao
DependenciaDestinoId
DependenciaDestinoDescricao
HoraPartida
HoraChegada
IsSuprimidoTotal
IsSuprimidoParcial
CargaRebocada
ComprimentoRebocado
IdMaterialMotor1
VelocidadeMaxima
Sentido
Comprimento

Circulação (Intermédio)
NumeroComboio1
NumeroComboio2
DataRealizacao
OperadorId
Operador
RegimeFrequenciaMnemonica
TipoServicoId
TipoServico
DataInicioCirculacao
DataFimCirculacao
DocumentoHorario
DependenciaOrigemId
DependenciaOrigemDescricao
DependenciaDestinoId
DependenciaDestinoDescricao
HoraPartida
HoraChegada
IsSuprimidoTotal
IsSuprimidoParcial
CargaRebocada
ComprimentoRebocado
IdMaterialMotor1
VelocidadeMaxima
Sentido
Comprimento
Comprimento Total
Peso Bruto
Carga Total

Circulação (Final)
VelocidadeMaxima
Comprimento Total
Carga Total

Anexo V

Ficheiros Finais Atributos (Pré-Integração)

G_Medições (Final)
ANO
ID_INSP
DIST_O
DIST_O_Trab
ID_POS
KM
LOCALIZ
PK
PKTrabalhado
P_SPEED
F_SPEED
SPEED
NIVLE
NIVLD
NIVLED1
NIVLDD1
NIVLED2
NIVLDD2
ALINE
ALIND
ALINED1
ALINDD1
ALINESQR
ALINDIRR
NIVTRANS
NIVTRANR
EMPENO3M
EMPREL3M
EMPENO9M
EMPREL9M
BITOLA
BITMED
GRAD
GRADMED
CURVA

Circulação (Final)
VelocidadeMaxima
Comprimento Total
Carga Total

Troço (Final)
PK
Inclinação
Curva de Concordância
Carril
Carril II
Travessa
Vel. Máxima de Planta

Anexo VI

Ficheiro Final

Antes da Eliminação Manual de Atributos

Ficheiro Final
ANO
ID_INSP
DIST_O
DIST_O_Trab
ID_POS
KM
LOCALIZ
PK
PKTrabalhado
Confirmação
P_SPEED
F_SPEED
SPEED
NIVLE
NIVLD
NIVLED1
NIVLDD1
NIVLED2
NIVLDD2
ALINE
ALIND
ALINED1
ALINDD1
ALINESQR
ALINDIRR
NIVTRANS
NIVTRANR
EMPENO3M
EMPREL3M
EMPENO9M
EMPREL9M
BITOLA
BITMED
GRAD
GRADMED
CURVA
Inclinação
Curva de Concordância
Carril
Carril II
Travessa
Vel. Máxima de Planta
Comprimento_Total (mediana)
Peso Total
Velocidade (Média)

Depois da Eliminação Manual de Atributos

Ficheiro Final
NIVLE
NIVLD
NIVLED1
NIVLDD1
NIVLED2
NIVLDD2
ALINE
ALIND
ALINED1
ALINDD1
ALINESQR
ALINDIRR
NIVTRANS
NIVTRANR
EMPENO3M
EMPREL3M
EMPENO9M
EMPREL9M
BITOLA
BITMED
GRAD
GRADMED
CURVA
Inclinação
Carril
Carril II
Travessa
Vel. Máxima de Planta
Comprimento_Total (mediana)
Peso Total
Velocidade (Média)

Anexo VII

Casos Omissos

```
     NIVLE             NIVLD            NIVLED1           NIVLDD1           NIVLED2           NIVLDD2           ALINE             ALIND             ALINED1
 Min.   :-27.3400   Min.   :-31.7600   Min.   :-17.3400   Min.   :-22.030   Min.   :-17.0300   Min.   :-16.8000   Min.   :-62.0300   Min.   :-63.7500   Min.   :-14.6500
 1st Qu.: -0.9800   1st Qu.: -0.9800   1st Qu.: -0.5900   1st Qu.: -0.590   1st Qu.: -1.3700   1st Qu.: -1.4100   1st Qu.: -5.9800   1st Qu.: -5.8600   1st Qu.: -0.7800
 Median :  0.0400   Median :  0.0400   Median :  0.0400   Median :  0.040   Median :  0.0000   Median :  0.0000   Median :  0.0400   Median :  0.0000   Median :  0.0000
 Mean   : -0.0001   Mean   : -0.0001   Mean   : -0.0008   Mean   : -0.001   Mean   :  0.0024   Mean   :  0.0022   Mean   :  0.1115   Mean   :  0.1117   Mean   : -0.0013
 3rd Qu.:  1.1300   3rd Qu.:  1.0900   3rd Qu.:  0.7000   3rd Qu.:  0.700   3rd Qu.:  1.3700   3rd Qu.:  1.4500   3rd Qu.:  6.4800   3rd Qu.:  6.4800   3rd Qu.:  0.7800
 Max.   : 18.7100   Max.   : 18.7500   Max.   : 13.9500   Max.   : 13.390   Max.   : 14.6100   Max.   : 17.2700   Max.   : 65.0000   Max.   : 64.6900   Max.   : 17.6600
 NA's   :1038       NA's   :1038       NA's   :1298       NA's   :1298      NA's   :1032       NA's   :1031       NA's   :800        NA's   :800        NA's   :1491
    ALINDD1           ALINESQR          ALINDIRR          NIVTRANS          NIVTRANR          EMPENO3M          EMPREL3M          EMPENO9M          EMPREL9M
 Min.   :-16.3300   Min.   :-27.5800   Min.   :-34.7300   Min.   :-197.970   Min.   :-19.1000   Min.   :-25.2000   Min.   :-24.2600   Min.   :-38.9800   Min.   :-12.9300
 1st Qu.: -0.8200   1st Qu.: -1.2100   1st Qu.: -1.2500   1st Qu.: -24.020   1st Qu.: -1.0900   1st Qu.: -1.4100   1st Qu.: -0.9000   1st Qu.: -2.4600   1st Qu.: -0.6200
 Median :  0.0000   Median :  0.0000   Median :  0.0000   Median :  -0.200   Median :  0.0000   Median :  0.0000   Median :  0.0000   Median :  0.0000   Median :  0.0000
 Mean   : -0.0012   Mean   :  0.0002   Mean   :  0.0004   Mean   :   1.103   Mean   : -0.0006   Mean   : -0.0005   Mean   :  0.0001   Mean   : -0.0015   Mean   :  0.0001
 3rd Qu.:  0.7800   3rd Qu.:  1.2100   3rd Qu.:  1.2500   3rd Qu.:  26.330   3rd Qu.:  1.0900   3rd Qu.:  1.4100   3rd Qu.:  0.9000   3rd Qu.:  2.4600   3rd Qu.:  0.6200
 Max.   : 15.2000   Max.   : 29.0200   Max.   : 25.7000   Max.   : 195.590   Max.   : 25.4700   Max.   : 21.0200   Max.   : 19.4900   Max.   : 40.2700   Max.   : 14.0200
 NA's   :1491       NA's   :1174       NA's   :1174       NA's   :611       NA's   :690        NA's   :617        NA's   :696        NA's   :629        NA's   :646
     BITOLA           BITMED            GRAD            GRADMED           CURVA                Inclinação      Curva.de.Concordância     Carril                Carril.II
 Min.   :-9.690   Min.   :-5.200   Min.   :-2.6600   Min.   :-2.4600   Min.   :-49.0200          : 69508    : 47578          Min.   :54.00   Barra Curta:138774
 1st Qu.: 0.230   1st Qu.: 0.550   1st Qu.: 0.3100   1st Qu.: 0.3100   1st Qu.: -5.4700   0,01268 :  7601   Não:180496       1st Qu.:54.00   BLS       :111230
 Median : 2.150   Median : 2.270   Median : 0.5500   Median : 0.5500   Median :  0.0000   0       :  6001   Sim: 21930       Median :54.00
 Mean   : 4.778   Mean   : 4.768   Mean   : 0.5231   Mean   : 0.5096   Mean   :  0.1016   -0,01629:  4000                    Mean   :56.31
 3rd Qu.: 8.360   3rd Qu.: 8.480   3rd Qu.: 0.8200   3rd Qu.: 0.7800   3rd Qu.:  6.2500   0,014673:  3608                    3rd Qu.:60.00
 Max.   :36.450   Max.   :27.430   Max.   : 2.5400   Max.   : 2.2300   Max.   : 52.6200   0,011658:  3392                    Max.   :60.00
 NA's   :611      NA's   :811      NA's   :611       NA's   :631       NA's   :691        (Other) :155894
        Travessa       Vel..Maxima.de.Planta Comprimento.Total..mediana.  Peso.Total      Velocidade.Média
 Betão Monobloco: 53684   Min.   : 75.0       Min.   :66.80              Min.   : 528482   Min.   : 67.0
 Bibloco        : 14861   1st Qu.: 90.0       1st Qu.:66.80              1st Qu.: 528482   1st Qu.:113.0
 Madeira        :138775   Median :110.0       Median :66.80              Median :1541135   Median :113.0
 Monobloco      : 42684   Mean   :107.1       Mean   :68.38              Mean   :2595228   Mean   :126.5
                          3rd Qu.:120.0       3rd Qu.:70.00              3rd Qu.:6271537   3rd Qu.:150.0
                          Max.   :140.0       Max.   :70.00              Max.   :7778428   Max.   :150.0
                          NA's   :47578
```

Anexo VIII

Caraterização Atributos Ficheiro Final

Ficheiro Final	Tipo de Variável	Escala
NIVLE	Numérica	Razão
NIVLD	Numérica	Razão
NIVLED1	Numérica	Razão
NIVLDD1	Numérica	Razão
NIVLED2	Numérica	Razão
NIVLDD2	Numérica	Razão
ALINE	Numérica	Razão
ALIND	Numérica	Razão
ALINED1	Numérica	Razão
ALINDD1	Numérica	Razão
ALINESQR	Numérica	Razão
ALINDIRR	Numérica	Razão
NIVTRANS	Numérica	Razão
NIVTRANR	Numérica	Razão
EMPENO3M	Numérica	Razão
EMPREL3M	Numérica	Razão
EMPENO9M	Numérica	Razão
EMPREL9M	Numérica	Razão
BITOLA	Numérica	Razão
BITMED	Numérica	Razão
GRAD	Numérica	Razão
GRADMED	Numérica	Razão
CURVA	Numérica	Razão
Inclinação	Numérica	Razão
Carril	Numérica	Razão
Carril II	Qualitativa	Nominal
Travessa	Qualitativa	Nominal
Vel. Máxima de Planta	Numérica	Razão
Comprimento_Total (mediana)	Numérica	Razão
Peso Total	Numérica	Razão
Velocidade (Média)	Numérica	Razão

Anexo IX

Caraterização Atributos Ficheiro Final por Antecedente (Caraterísticas da Via) e Consequente (Parâmetros de Via)

Ficheiro Final	Tipo de Variável	Escala
NIVLE	Numérica	Razão
NIVLD	Numérica	Razão
NIVLED1	Numérica	Razão
NIVLDD1	Numérica	Razão
NIVLED2	Numérica	Razão
NIVLDD2	Numérica	Razão
ALINE	Numérica	Razão
ALIND	Numérica	Razão
ALINED1	Numérica	Razão
ALINDD1	Numérica	Razão
ALINESQR	Numérica	Razão
ALINDIRR	Numérica	Razão
NIVTRANS	Numérica	Razão
NIVTRANR	Numérica	Razão
EMPENO3M	Numérica	Razão
EMPREL3M	Numérica	Razão
EMPENO9M	Numérica	Razão
EMPREL9M	Numérica	Razão
BITOLA	Numérica	Razão
BITMED	Numérica	Razão
GRAD	Numérica	Razão
GRADMED	Numérica	Razão
CURVA	Numérica	Razão
Inclinação	Numérica	Razão
Carril	Numéria	Razão
Carril II	Qualitativa	Nominal
Travessa	Qualitativa	Nominal
Vel. Máxima de Planta	Numérica	Razão
Comprimento_Total (mediana)	Numérica	Razão
Peso Total	Numérica	Razão
Velocidade (Média)	Numérica	Razão

Anexo X

Ficheiro Final (Atributos Caracterizadores da Via)	Tipo de Variável	Escala	Potencial Informativo (*)
GRAD	Numérica	Razão	Informativo
GRADMED	Numérica	Razão	Informativo
CURVA	Numérica	Razão	Informativo
Inclinação	Numérica	Razão	Informativo
Carril II	Qualitativa	Nominal	Pouco informativo
Travessa	Qualitativa	Nominal	Pouco informativo
Carril	Numéria	Razão	Pouco informativo
Vel. Máxima de Placa	Numérica	Razão	Pouco informativo
Comprimento_Total (mediana)	Numérica	Razão	Pouco informativo
Peso Total	Numérica	Razão	Pouco informativo
Velocidade (Média)	Numérica	Razão	Pouco informativo

(*) - Com base na diversidade de valores que cada atributo apresenta

CarrilII

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Bar	86234	48,2	48,2	48,2
	BLS	92681	51,8	51,8	100,0
	Total	178915	100,0	100,0	

Carril

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	54	100384	56,1	56,1	56,1
	60	78531	43,9	43,9	100,0
	Total	178915	100,0	100,0	

Travessa

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Betão M	78531	43,9	43,9	43,9
	Bibloco	14149	7,9	7,9	51,8
	Madeira	86235	48,2	48,2	100,0
	Total	178915	100,0	100,0	

Vel.Máxima de Placa

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	67	14149	7,9	7,9	7,9
	113	73027	40,8	40,8	48,7
	146	13207	7,4	7,4	56,1
	150	78532	43,9	43,9	100,0
	Total	178915	100,0	100,0	

Comprimento_Total

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	66,8	105888	59,2	59,2	59,2
	70,8	73027	40,8	40,8	100,0
	Total	178915	100,0	100,0	

Peso.Total

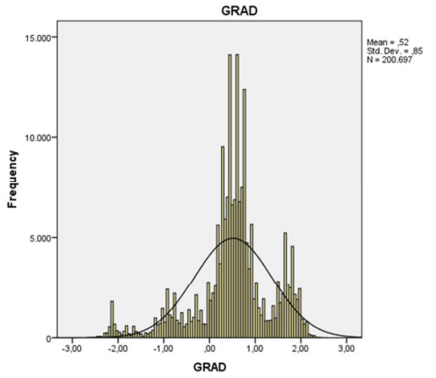
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	329448,0	73027	40,8	40,8	40,8
	1541998,0	13207	7,4	7,4	48,2
	1623998,0	92681	51,8	51,8	100,0
	1953448,0	78531	43,9	43,9	43,9
	2277998,0	14149	7,9	7,9	51,8
	Total	178915	100,0	100,0	

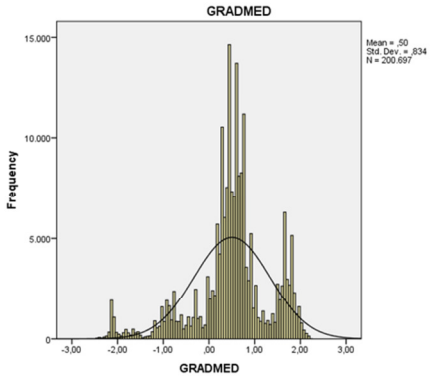
Velocidade

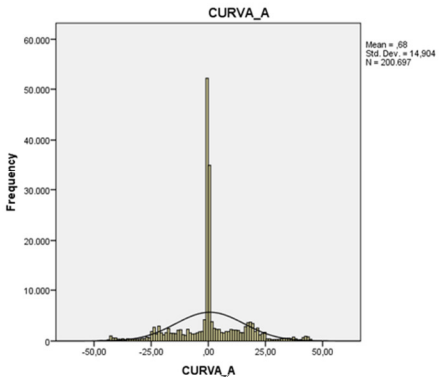
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	67	14149	7,9	7,9	7,9
	113	73027	40,8	40,8	48,7
	146	13207	7,4	7,4	56,1
	150	78532	43,9	43,9	100,0
	Total	178915	100,0	100,0	

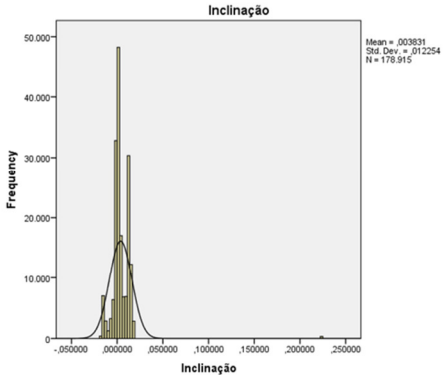
Anexo XI

Análise Univariada – Variáveis a Discretizar









Statistics

		GRAD	GRADMED	CURVA_A	Inclinação
N	Valid	200697	200697	200697	178915
	Missing	0	0	0	21782
Mean		,5198	,5042	,6840	,003831047
Median		,5500	,5500	,0000	,001982100
Std. Deviation		,84978	,83431	14,90408	,012254856
Skewness		-,346	-,337	,001	,502
Std. Error of Skewness		,005	,005	,005	,006
Kurtosis		1,209	1,227	,000	1,452
Std. Error of Kurtosis		,011	,011	,011	,012
Range		5,20	5,00	99,93	,241576

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
GRAD	,144	178915	,000
GRADMED	,146	178915	,000
CURVA_A	,167	178915	,000
Inclinação	,167	178915	,000

a. Lilliefors Significance Correction

Anexo XII

Teste KS à Normalidade das Distribuições

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
NIVLE	,098	200697	,000
NIVLD	,104	200697	,000
NIVLED1	,109	200697	,000
NIVLDD1	,119	200697	,000
NIVLED2	,057	200697	,000
NIVLDD2	,056	200697	,000
ALINE	,153	200697	,000
ALIND	,152	200697	,000
ALINED1	,088	200697	,000
ALINDD1	,082	200697	,000
ALINESQR	,093	200697	,000
ALINDIRR	,086	200697	,000
NIVTRANS	,198	200697	,000
NIVTRANR	,084	200697	,000
EMPENO3M	,088	200697	,000
EMPREL3M	,095	200697	,000
EMPENO9M	,129	200697	,000
EMPREL9M	,093	200697	,000
BITOLA	,192	200697	,000
BITMED	,193	200697	,000

a. Lilliefors Significance Correction

Anexo XIII

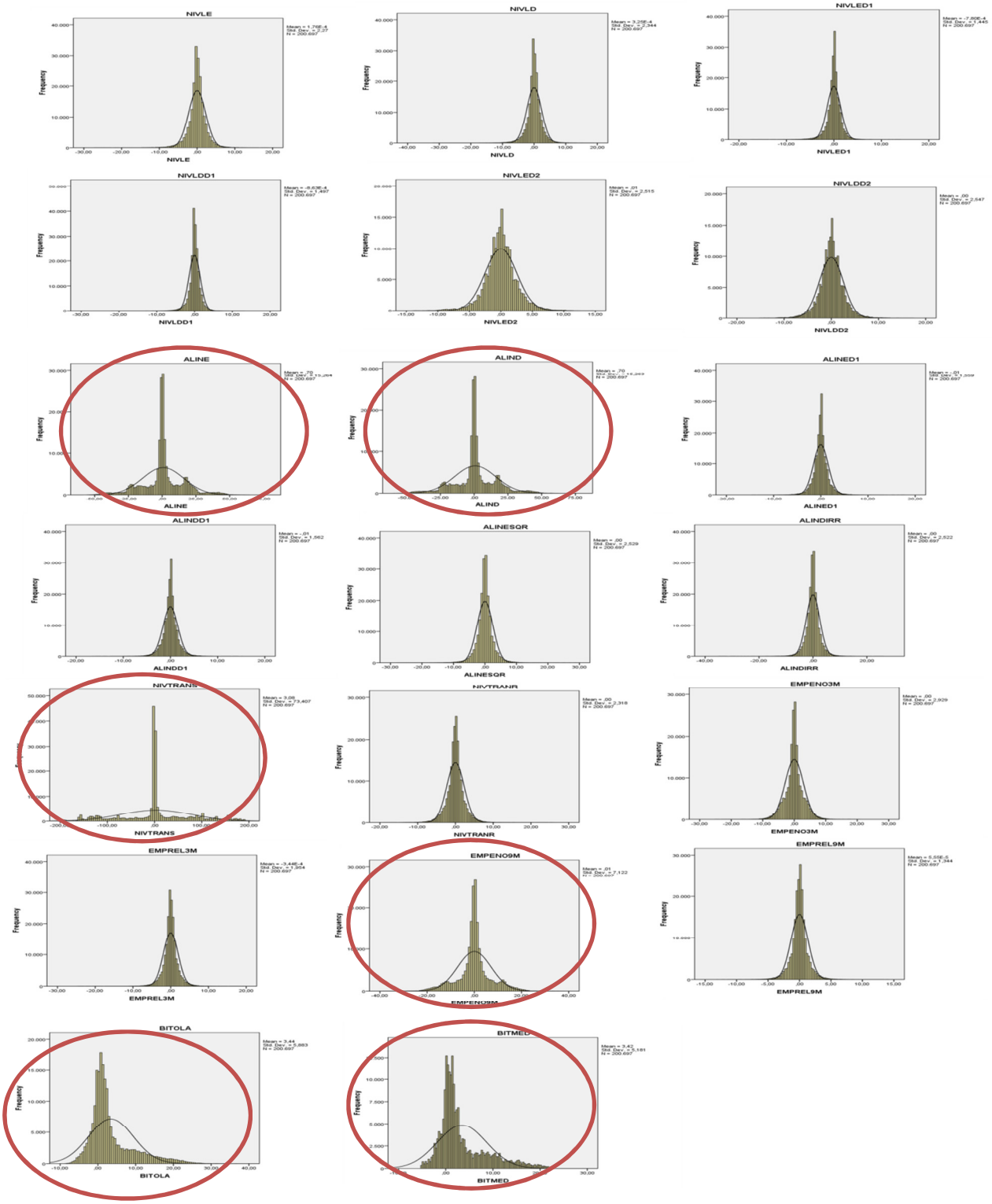
Estatística Descritiva Atributos Consequentes

Statistics

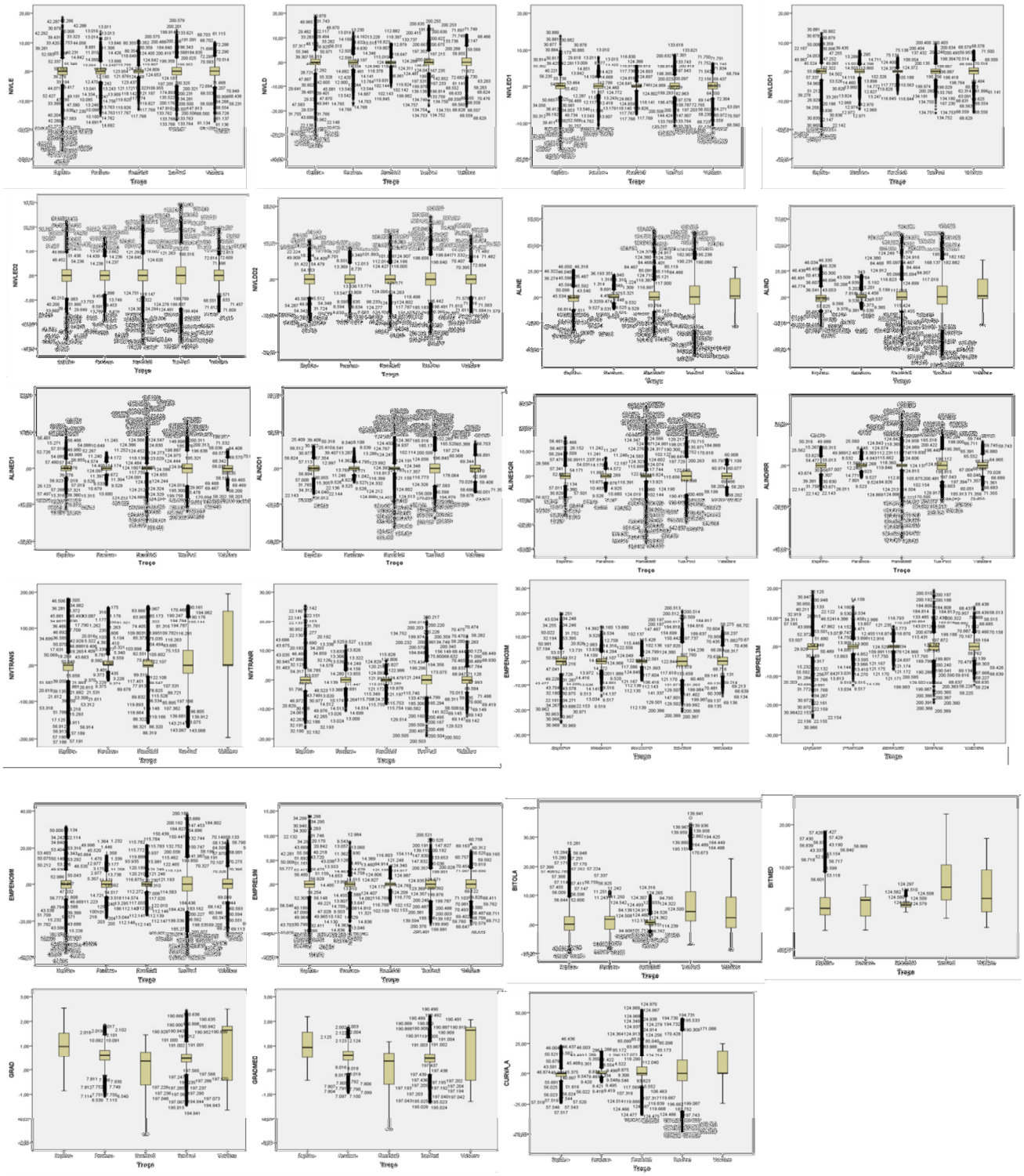
		NIVLE	NIVLD	NIVLED1	NIVLDD1	NIVLED2	NIVLDD2	ALINE	ALIND	ALINED1	ALINDD1
N	Valid	200697	200697	200697	200697	200697	200697	200697	200697	200697	200697
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		,0002	,0003	-,0008	-,0009	,0056	,0047	,6968	,6968	-,0064	-0,0064
Median		,0400	,0400	,0400	,0400	0,0000	0,0000	,1600	,1600	0,0000	0,0000
Std. Deviation		2,27016	2,34359	1,44540	1,49662	2,51538	2,54653	15,26450	15,26325	1,55861	1,56177
Skewness		-,665	-,824	-,949	-1,153	-,120	-,073	,001	-,004	,006	-0,027
Std. Error of		,005	,005	,005	,005	,005	,005	,005	,005	,005	,005
Kurtosis		7,427	10,580	8,179	11,848	2,592	2,718	1,442	1,449	6,025	5,356
Std. Error of Kurtosis		,011	,011	,011	,011	,011	,011	,011	,011	,011	,011
Range		46,05	50,51	31,29	35,35	28,52	32,31	120,51	120,82	32,31	31,53

		ALINESQR	ALINDIRR	NIVTRANS	NIVTRANR	EMPENO3M	EMPREL3M	EMPENO9M	EMPREL9M	BITOLA	BITMED
N	Valid	200697	200697	200697	200697	200697	200697	200697	200697	200697	200697
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		,0016	,0017	3,0770	,0017	,0018	-,0003	,0055	,0001	3,4376	3,4239
Median		0,0000	0,0000	,0400	0,0000	0,0000	0,0000	0,0000	0,0000	1,5600	1,6000
Std. Deviation		2,52906	2,52182	73,40732	2,31832	2,92921	1,95393	7,12159	1,34437	5,88274	5,18141
Skewness		-,011	-,165	,026	,181	-,052	-,048	-,026	-,008	1,492	1,361
Std. Error of		,005	,005	,005	,005	,005	,005	,005	,005	,005	,005
Kurtosis		8,431	8,795	,438	5,773	2,653	8,565	1,863	6,077	2,100	1,384
Std. Error of Kurtosis		,011	,011	,011	,011	,011	,011	,011	,011	,011	,011
Range		56,60	60,43	393,56	44,57	46,22	43,75	76,95	26,95	46,14	27,74

Anexo XIV


NIVLE
NIVLD
NIVLED1
NIVLDD1
NIVLED2
NIVLDD2
ALINE
ALIND
ALINED1
ALINDD1
ALINESQR
ALINDIRR
NIVTRANS
NIVTRANR
EMPENO3M
EMPREL3M
EMPENO9M
EMPREL9M
BITOLA
BITMED
Frequency

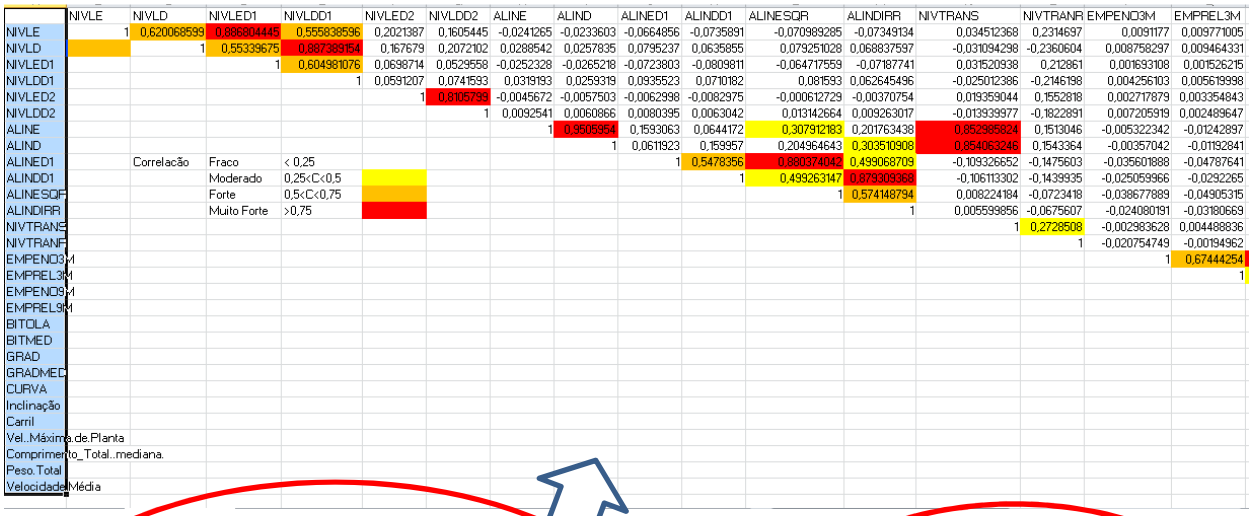
Anexo XV


NIVLE
NIVLD
NIVLED1
NIVLDD1
NIVLED2
NIVLDD2
ALINE
ALIND
ALINED1
ALINDD1
ALINESOR
ALINDRR
NIVTRANS
NIVTRANR
EMPENOM
EMPRELJM
EMPENOM
EMPRELJN
BITOLA
BITMED
GRAD
GRACMED
CURVA_A
Espinho
Paramos
Mandaetti
Tua Prot
Valdaire
Troço

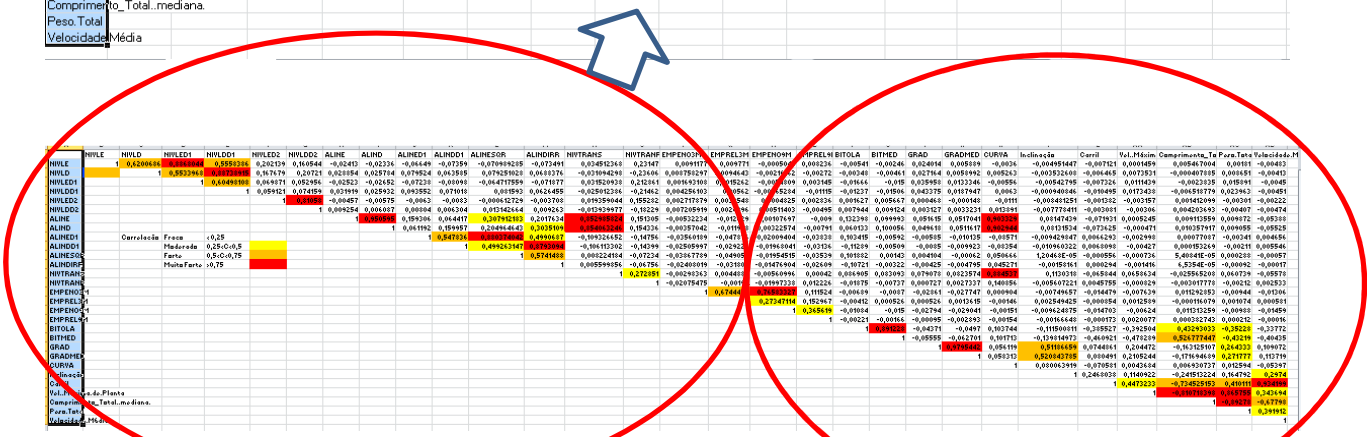
Anexo XVI

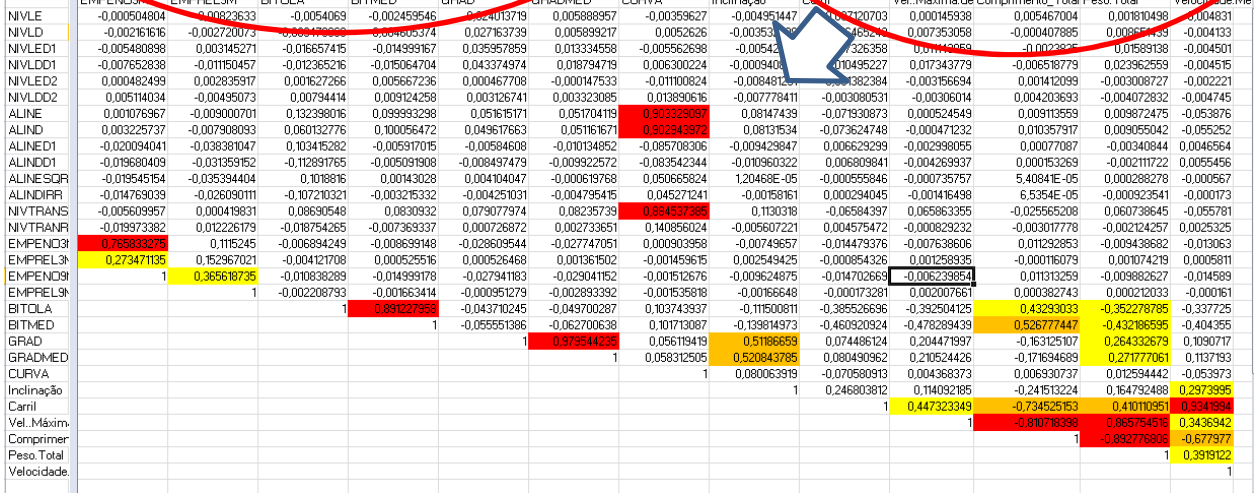
Análise Bivariada

Correlação Spearman



	NIVLE	NIVLD	NIVLED1	NIVLDD1	NIVLED2	NIVLDD2	ALINE	ALIND	ALINED1	ALINDD1	ALINESQR	ALINDIRR	NIVTRANS	NIVTRANR	EMPENO3M	EMPREL3M
NIVLE	1	0,620068599	0,886804445	0,555838596	0,2021387	0,1605445	-0,0241265	-0,0233603	-0,0664856	-0,0735891	-0,070989285	-0,07349134	0,034512368	0,2314697	0,0091177	0,009771005
NIVLD		1	0,55339675	0,887389154	0,167679	0,2072102	0,0288542	0,0257835	0,0795237	0,0635855	0,079251028	0,068837597	-0,031094298	-0,2360604	0,008758297	0,009464331
NIVLED1			1	0,604981076	0,0698714	0,0529558	-0,0252328	-0,0265218	-0,0723803	-0,0809811	-0,064717559	-0,07187741	0,031520938	0,212861	0,001693108	0,001526215
NIVLDD1				1	0,0591207	0,0741593	0,0319193	0,0259319	0,0935523	0,0710182	0,081593	0,062645496	-0,025012386	-0,2146198	0,004256103	0,005619998
NIVLED2					1	0,8105793	-0,0045672	-0,0057503	-0,0062998	-0,0082975	-0,000612729	-0,00370754	0,019359044	0,1552818	0,002717879	0,000354843
NIVLDD2						1	0,0092541	0,0060866	0,0080395	0,0063042	0,013142664	0,009263017	-0,013939977	-0,1822891	0,007205919	0,002489647
ALINE							1	0,9505954	0,1593063	0,0644172	0,307912183	0,201763438	0,852985824	0,1513046	-0,005322342	-0,01242897
ALIND								1	0,0611923	0,159957	0,204964643	0,303510908	0,854063246	0,1543364	-0,00357042	-0,01192841
ALINED1		Correlação	Fraco	< 0,25					1	0,5478356	0,880374042	0,499068709	-0,109326652	-0,1475603	-0,035601888	-0,04787641
ALINDD1			Moderado	0,25<C<0,5						1	0,499263147	0,878309368	-0,106113302	-0,1439935	-0,025059966	-0,0292265
ALINESQR			Forte	0,5<C<0,75							1	0,574148794	0,008224184	-0,0723418	-0,038677889	-0,04905315
ALINDIRR			Muito Forte	>0,75								1	0,005599856	-0,0675607	-0,024080191	-0,03180669
NIVTRANS													1	0,2728508	-0,002983628	0,004488836
NIVTRANR														1	-0,020754749	-0,00194962
EMPENO3M															1	0,67444254
EMPREL3M																1
EMPENO9M																
EMPREL9M																
BITOLA																
BITMED																
GRAD																
GRADMED																
CURVA																
Inclinação																
Carril																
Vel..Máxima.de.Planta																
Comprimento_Total..mediana.																
Peso.Total																
Velocidade.Média																

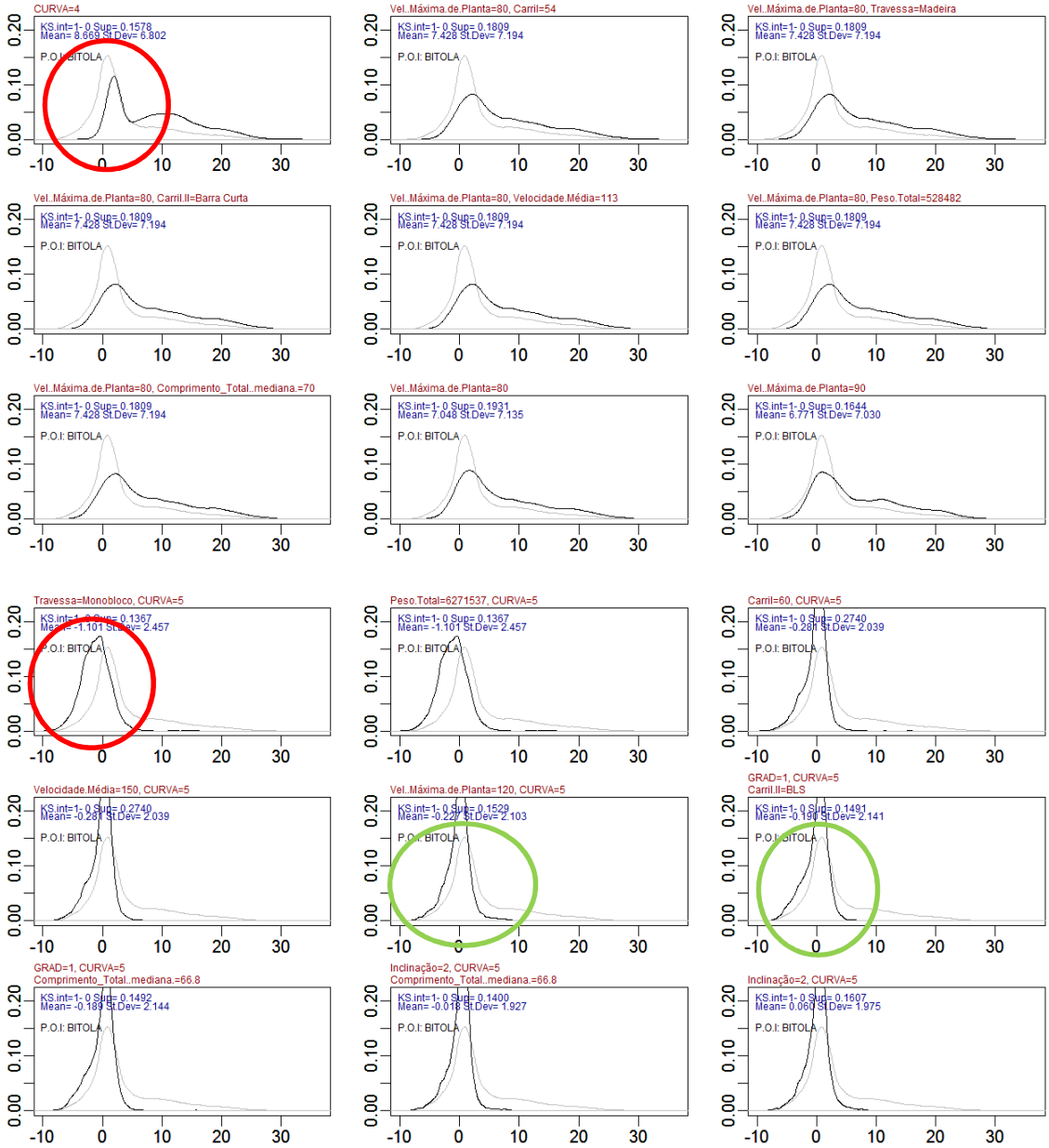


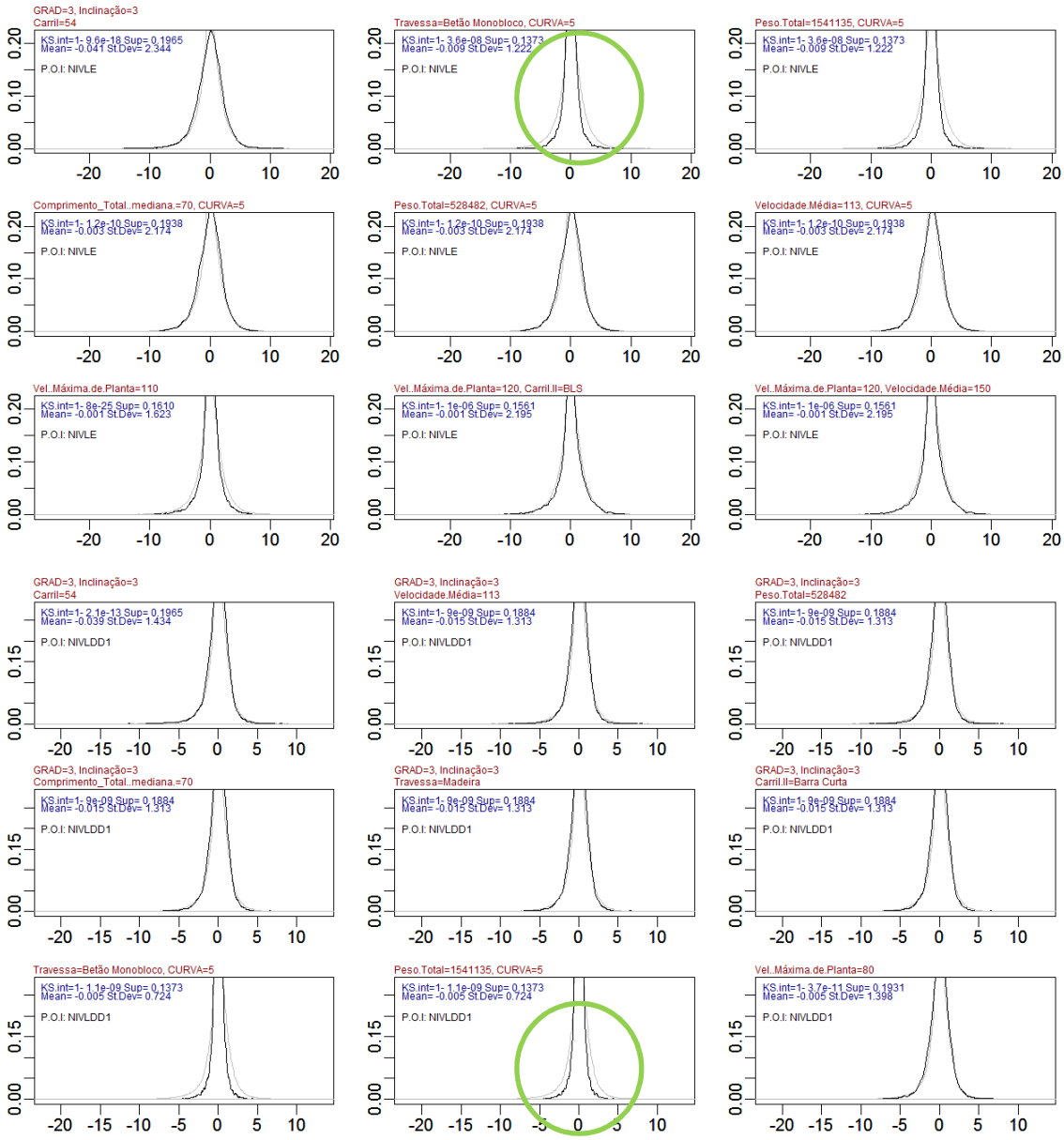


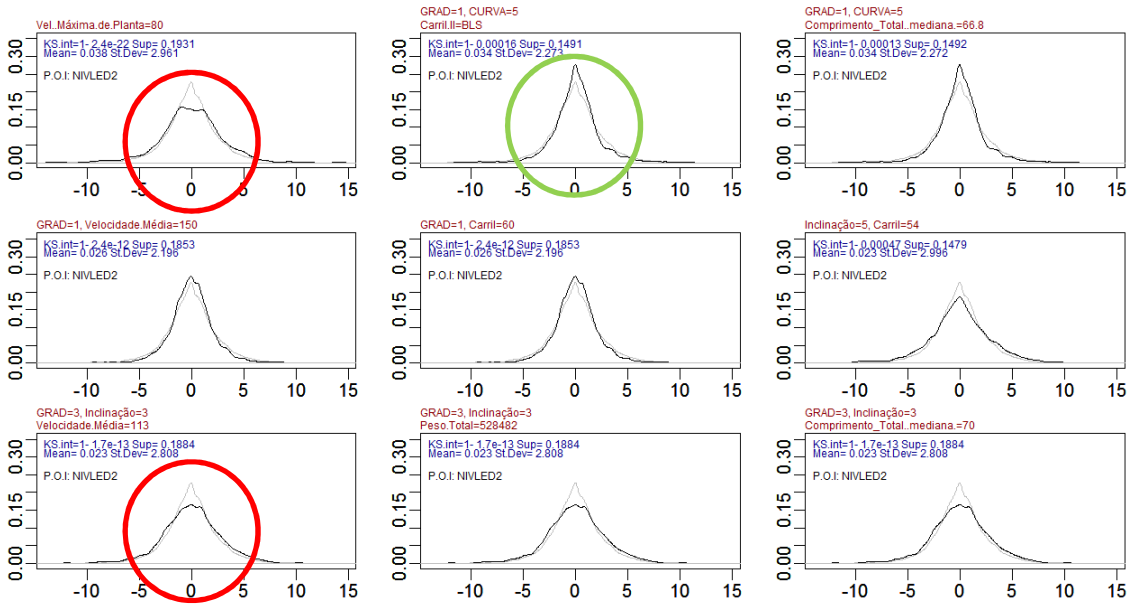
	EMPENO3M	EMPREL3M	BITOLA	BITMED	GRAD	GRADMED	CURVA	Inclinação	Carril	Vel..Máxima.de.Planta	Comprimento_Total	Peso.Total	Velocidade.Mé
NIVLE	-0,000504804	0,00823633	-0,0054069	-0,002459546	0,024013719	0,005888957	-0,00359627	-0,004951447	-0,00720703	0,000145938	0,005467004	0,001810498	0,004831
NIVLD	-0,002161616	-0,002720073	-0,0084478868	-0,004605374	0,027163739	0,005899217	0,0052626	-0,00353	-0,0052249	0,007353058	-0,000407885	0,00855499	-0,004133
NIVLED1	-0,005480898	0,003145271	-0,016657415	-0,014999167	0,035957859	0,013334558	-0,005562698	-0,00542	0,00326358	0,01119059	-0,0023835	0,01589138	-0,004501
NIVLDD1	-0,007652838	-0,011150457	-0,012365216	-0,015064704	0,043374974	0,018794719	0,006300224	-0,0009408	0,010495227	0,017343779	-0,006518779	0,023962559	-0,004515
NIVLED2	0,000482499	0,002835917	0,001627266	0,005667236	0,000467708	-0,000147533	-0,011100824	-0,0084812	0,00382384	-0,003156694	0,001412099	-0,003008727	-0,002221
NIVLDD2	0,005114034	-0,0049507	0,00794414	0,009124258	0,003126741	0,003323085	0,013890616	-0,007778411	-0,003080531	-0,00306014	0,004203693	-0,004072832	-0,004745
ALINE	0,001076967	-0,009000701	0,132398016	0,099393298	0,051615171	0,051704119	0,903329097	0,08147439	-0,071930873	0,000524549	0,009113559	0,009872475	-0,053876
ALIND	0,003225737	-0,007908093	0,060132776	0,100056472	0,049617663	0,051161671	0,902943972	0,08131534	-0,073624748	-0,000471232	0,010357917	0,009055042	-0,055252
ALINED1	-0,020094041	-0,038381047	0,103415282	-0,005917015	-0,00584608	-0,010134852	-0,085708306	-0,009429847	0,006629299	-0,002998055	0,00077087	-0,00340844	0,0046564
ALINDD1	-0,019680409	-0,031359152	-0,112891765	-0,005091908	-0,008497479	-0,009922572	-0,083542344	-0,010960322	0,006809841	-0,004269937	0,000153269	-0,002111722	0,0055456
ALINESQR	-0,019545154	-0,035394404	0,1018816	0,00143028	0,004104047	-0,000619768	0,050665824	1,20468E-05	-0,000555846	-0,000735757	5,40841E-05	0,000288278	-0,000567
ALINDIRR	-0,014769039	-0,026090111	-0,107210321	-0,003215332	-0,004251031	-0,004795415	0,045271241	-0,00158161	0,000294045	-0,001416498	6,5354E-05	-0,000923541	-0,000173
NIVTRANS	-0,005609957	0,000419831	0,08690548	0,0830932	0,079077974	0,08235739	0,884537385	0,1130318	-0,06584397	0,065863355	-0,025565208	0,060738645	-0,055781
NIVTRANR	-0,019973382	0,012226179	-0,018754265	-0,007369337	0,000726872	0,002733651	0,140856024	-0,005607221	0,004575472	-0,000829232	-0,003017778	-0,002124257	0,0025325
EMPENO3M	0,765833275	0,1115245	-0,006894249	-0,008699148	-0,028609544	-0,027747051	0,000903958	-0,00749657	-0,014479376	-0,007638606	0,011292853	-0,009438682	-0,013063
EMPREL3M	0,273471135	0,152967021	-0,004121708	0,000525516	0,000526468	0,001361502	-0,001459615	0,002549425	-0,000854326	0,001258935	-0,000116079	0,001074219	0,0005811
EMPENO9M	1	0,365618735	-0,010838289	-0,014999178	-0,027941183	-0,029041152	-0,001512676	-0,009624875	-0,014702669	-0,006239854	0,011313259	-0,009882627	-0,014589
EMPREL9M		1	-0,002208793	-0,001663414	-0,000951279	-0,002893392	-0,001535818	-0,00166648	-0,000173281	0,002007661	0,000382743	0,000212033	-0,000161
BITOLA			1	0,891227958	-0,043710245	-0,049700287	0,103743937	-0,111500811	-0,385526696	-0,392504125	0,43293033	-0,352278785	-0,337725
BITMED				1	-0,055551386	-0,062700638	0,101713087	-0,139814973	-0,460920924	-0,478289439	0,526777447	-0,432186595	-0,404355
GRAD					1	0,979544235	0,056119419	0,51186659	0,074486124	0,204471997	-0,163125107	0,264332679	0,1090717
GRADMED						1	0,058312505	0,520843785	0,080490962	0,210524426	-0,171694689	0,271777061	0,1137193
CURVA							1	0,080063919	-0,070580913	0,004368373	0,006930737	0,012594442	-0,053973
Inclinação								1	0,246803812	0,114092185	-0,241513224	0,164792488	0,2973995
Carril									1	0,447323349	-0,734525153	0,410110951	0,9341994
Vel..Máxim										1	-0,810719398	0,865754516	0,3436942
Comprimer											1	-0,892776806	-0,677977
Peso.Total												1	0,3919122
Velocidade.													1

Anexo XVII

Distribution Rules




GRAD=3, Inclinação=3
Carril=54
KS.int=1- 9.6e-18 Sup= 0.1965
Mean= -0.041 St.Dev= 2.344
P.O.I: NIVLE
Travessa=Betão Monobloco, CURVA=5
KS.int=1- 3.6e-08 Sup= 0.1373
Mean= -0.009 St.Dev= 1.222
P.O.I: NIVLE
Peso.Total=1541135, CURVA=5
KS.int=1- 3.6e-08 Sup= 0.1373
Mean= -0.009 St.Dev= 1.222
P.O.I: NIVLE
Comprimento_Total..mediana.=70, CURVA=5
KS.int=1- 1.2e-10 Sup= 0.1938
Mean= -0.003 St.Dev= 2.174
P.O.I: NIVLE
Peso.Total=528482, CURVA=5
KS.int=1- 1.2e-10 Sup= 0.1938
Mean= -0.003 St.Dev= 2.174
P.O.I: NIVLE
Velocidade.Média=113, CURVA=5
KS.int=1- 1.2e-10 Sup= 0.1938
Mean= -0.003 St.Dev= 2.174
P.O.I: NIVLE
Vel..Máxima.de.Planta=110
KS.int=1- 8e-25 Sup= 0.1610
Mean= -0.001 St.Dev= 1.823
P.O.I: NIVLE
Vel..Máxima.de.Planta=120, Carril.II=BLS
KS.int=1- 1e-06 Sup= 0.1561
Mean= -0.001 St.Dev= 2.195
P.O.I: NIVLE
Vel..Máxima.de.Planta=120, Velocidade.Média=150
KS.int=1- 1e-06 Sup= 0.1561
Mean= -0.001 St.Dev= 2.195
P.O.I: NIVLE
GRAD=3, Inclinação=3
Carril=54
KS.int=1- 2.1e-13 Sup= 0.1965
Mean= -0.039 St.Dev= 1.434
P.O.I: NIVLDD1
GRAD=3, Inclinação=3
Velocidade.Média=113
KS.int=1- 9e-09 Sup= 0.1884
Mean= -0.015 St.Dev= 1.313
P.O.I: NIVLDD1
GRAD=3, Inclinação=3
Peso.Total=528482
KS.int=1- 9e-09 Sup= 0.1884
Mean= -0.015 St.Dev= 1.313
P.O.I: NIVLDD1
GRAD=3, Inclinação=3
Comprimento_Total..mediana.=70
KS.int=1- 9e-09 Sup= 0.1884
Mean= -0.015 St.Dev= 1.313
P.O.I: NIVLDD1
GRAD=3, Inclinação=3
Travessa=Madeira
KS.int=1- 9e-09 Sup= 0.1884
Mean= -0.015 St.Dev= 1.313
P.O.I: NIVLDD1
GRAD=3, Inclinação=3
Carril.II=Barra Curta
KS.int=1- 9e-09 Sup= 0.1884
Mean= -0.015 St.Dev= 1.313
P.O.I: NIVLDD1
Travessa=Betão Monobloco, CURVA=5
KS.int=1- 1.1e-09 Sup= 0.1373
Mean= -0.005 St.Dev= 0.724
P.O.I: NIVLDD1
Peso.Total=1541135, CURVA=5
KS.int=1- 1.1e-09 Sup= 0.1373
Mean= -0.005 St.Dev= 0.724
P.O.I: NIVLDD1
Vel..Máxima.de.Planta=80
KS.int=1- 3.7e-11 Sup= 0.1931
Mean= -0.005 St.Dev= 1.398
P.O.I: NIVLDD1

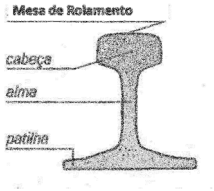

Vel..Máxima.de.Planta=80
KS.int=1- 2.4e-22 Sup= 0.1931
Mean= 0.038 St.Dev= 2.961
P.O.I: NIVLED2
GRAD=1, CURVA=5
Carril.II=BLS
KS.int=1- 0.00016 Sup= 0.1491
Mean= 0.034 St.Dev= 2.273
P.O.I: NIVLED2
GRAD=1, CURVA=5
Comprimento_Total..mediana.=66.8
KS.int=1- 0.00013 Sup= 0.1492
Mean= 0.034 St.Dev= 2.272
P.O.I: NIVLED2
GRAD=1, Velocidade.Média=150
KS.int=1- 2.4e-12 Sup= 0.1853
Mean= 0.026 St.Dev= 2.196
P.O.I: NIVLED2
GRAD=1, Carril=60
KS.int=1- 2.4e-12 Sup= 0.1853
Mean= 0.026 St.Dev= 2.196
P.O.I: NIVLED2
Inclinação=5, Carril=54
KS.int=1- 0.00047 Sup= 0.1479
Mean= 0.023 St.Dev= 2.996
P.O.I: NIVLED2
GRAD=3, Inclinação=3
Velocidade.Média=113
KS.int=1- 1.7e-13 Sup= 0.1884
Mean= 0.023 St.Dev= 2.808
P.O.I: NIVLED2
GRAD=3, Inclinação=3
Peso.Total=528482
KS.int=1- 1.7e-13 Sup= 0.1884
Mean= 0.023 St.Dev= 2.808
P.O.I: NIVLED2
GRAD=3, Inclinação=3
Comprimento_Total..mediana.=70
KS.int=1- 1.7e-13 Sup= 0.1884
Mean= 0.023 St.Dev= 2.808
P.O.I: NIVLED2
0.00
0.15
0.30
-10
-5
0
5
10
15

Anexo XVIII

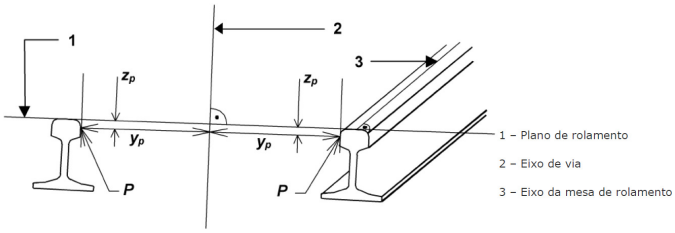
Glossário

Este glossário visa documentar alguns conceitos associados ao domínio em estudo

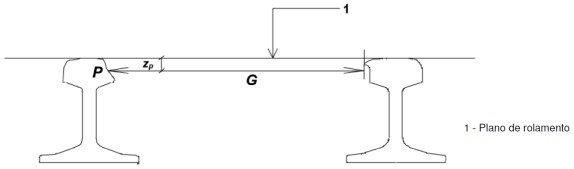
Carril ► Barra de ferro, fixa em travessas de madeira ou de betão, e sobre a qual se movem as rodas de diferentes veículos.



Alinhamento ► Em termos comuns significa o desvio, expresso por y_p de cada fila de carril, medido em diversas posições de p face a uma linha intermédia que representa o eixo da via. Esta métrica ….



Bitola (Pontual) ► Em termos comuns é o afastamento entre carris, medido pela menor distância G entre as faces internas da cabeça de dois carris adjacentes é medido no ponto P a uma distância Z_p do plano de rolamento com uma variação pré definida.



Bitola Média ► Representa a média, em 100 metros, das observações da bitola pontual.

Empeno ► Definidos quatro pontos sobre a mesa de rolamento dos carris, o que corresponde a dois pontos sobre cada um dos carris, empeno define-se como a distância vertical de um dos pontos ao plano formado pelos outros três.

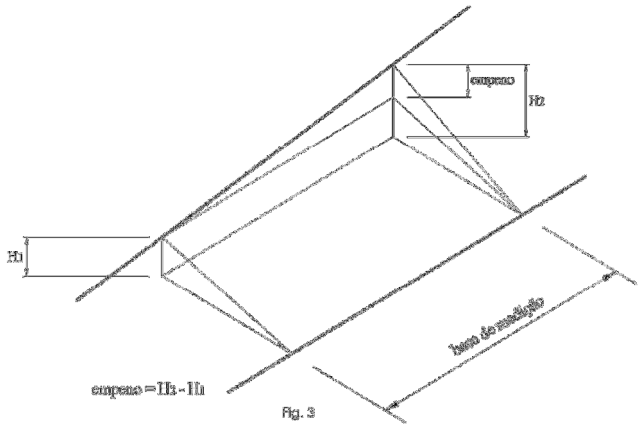
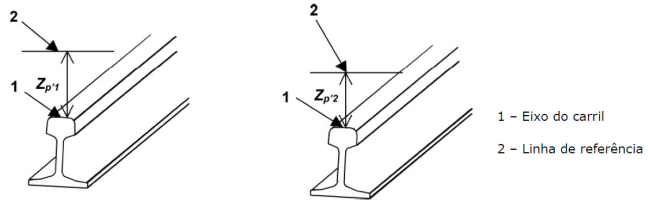


Fig. 3

Nivelamento Longitudinal ► Em termos comuns é aferir se o carril se encontra ao nível face a um eixo de referência teórico ou seja, se verticalmente está abaixo ou acima do que seria suposto. Esta medição é efetuada no eixo do carril e é dada pelo desvio Z_p relativamente a uma linha de referência perpendicular ao plano de rolamento.



Nivelamento Transversal ► Em termos comuns é a diferença de altura entre dois carris, aferida na mesa de rolamento de cada carril topo do carril, face a um padrão de referência calculado na forma infra por:

1. Nivelamento transversal
2. Plano de rolamento
3. Plano horizontal de referência
4. Distância entre eixos dos carris.